МІНІСТЕРСТВО ОСВІТИ І НАУКИ,
МОЛОДІ ТА СПОРТУ УКРАЇНИ

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
"ХАРКІВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ"

Матеріали

І Всеукраїнської науково-практичної конференції

"ІНТЕЛЕКТУАЛЬНІ СИСТЕМИ ТА ПРИКЛАДНА ЛІНГВІСТИКА"

**ПРИСВЯЧУЄТЬСЯ П'ЯТИРІЧЧЮ
КАФЕДРИ ІНТЕЛЕКТУАЛЬНИХ
КОМП'ЮТЕРНИХ СИСТЕМ**

ХАРКІВ 2012

УДК 004.9:81

Матеріали І Всеукраїнської науково-практичної конференції "Інтелектуальні системи та прикладна лінгвістика".

Харків, 15-16 березня 2012 р.: Тези доповідей. – Харків: Національний технічний університет "Харківський політехнічний інститут", 2012. – 78 с.

В матеріалах розглядаються проблеми та перспективи розвитку інтелектуальних комп’ютерних систем та різних галузей прикладної лінгвістики, а саме корпусної лінгвістики, комп'ютерної лексикографії, машинного перекладу, лінгвістики Інтернету; питання використання інформаційних технологій в лінгвістиці, з метою дослідження та обробки мови.

ЗМІСТ

Шаронова Н.В.
ПРИКЛАДНОЙ ЛИНГВИСТИКЕ В НТУ «ХПИ» И КАФЕДРЕ ИНТЕЛЛЕКТУАЛЬНЫХ КОМПЬЮТЕРНЫХ СИСТЕМ – ПЯТЬ ЛЕТ! 6

Дорошенко А.Ю.
ВИКОРИСТАННЯ ОНТОЛОГІЙ ДЛЯ СЕМАНТИЧНОГО ПОШУКУ ДОКУМЕНТІВ…… 8

Говорущенко Т.О.
НЕЙРОМЕРЕЖНА СИСТЕМА ВАЛІДАЦІЇ ПРОЕКТІВ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ…… 10

Orobinska O.O.
GENERALIZING FRAMEWORK FOR ONTOLOGY LEARNING……12

Неручок Ю. Ю.
МОДЕЛИРОВАНИЕ ОСНОВНЫХ ПОЛОЖЕНИЙ АППРОКСИМИРОВАННОЙ СТРУКТУРЫ СОДЕРЖАНИЯ ХУДОЖЕСТВЕННОГО ПРОИЗВЕДЕНИЯ…… 14

Іщенко О.С.
ЕКСПЕРИМЕНТАЛЬНА ФОНЕТИКА У СВІТЛІ СУЧАСНИХ КОМП'ЮТЕРНИХ ТЕХНОЛОГІЙ…… 16

Михайлов М.С.
СТВОРЕННЯ ТЕСТІВ З ГРАМАТИКИ АНГЛІЙСЬКОЇ МОВИ ЗА ГРАМАТИЧНИМИ ТЕМАМИ«REPERTED SPEECH» ТА «SEQUENCES OF TENSES»…….. 18

Кочуєва З.А.
МЕТОДИ І МОДЕЛІ ОБРОБКИ ІНФОРМАЦІЙНИХ ОБ'ЄКТІВ У СУЧАСНИХ БІБЛІОТЕЧНИХ СИСТЕМАХ……...20

Агеев И.В.
РЕАЛИЗАЦИЯ ВЫЯВЛЕНИЯ И ИСПРАВЛЕНИЯ ОРФОГРАФИЧЕСКИХ ОШИБОК В ТЕКСТЕ…… 22

Гостева Е.С.
ИЗВЛЕЧЕНИЕ КЛЮЧЕВЫХ ПОНЯТИЙ ИЗ ТЕКСТОВЫХ ДОКУМЕНТОВ ПРИ ГЕНЕРАЦИИ ПРЕСС-ПОРТРЕТА…… 24

Дашкевич Е.С.
ПОСТРОЕНИЕ И ИСПОЛЬЗОВАНИЕ МАТРИЦЫ ИНЦИДЕНТНОСТИ «ТЕРМИН-ДОКУМЕНТ» В ПОЛНОТЕКСТОВОМ ИНФОРМАЦИОННОМ ПОИСКЕ…….. 26

Лебедєва Г.М.
ЛЕКСИЧНО-СЕМАНТИЧНІ ОСОБЛИВОСТІ ТА ОСОБЛИВОСТІ ПЕРЕКЛАДУ ОКАЗІОНАЛІЗМІВ (НА МАТЕРІАЛІ РОМАНУ ДЖ. РОУЛІНГ "ГАРРІ ПОТТЕР І ОРДЕН ФЕНІКСУ")…….. 28

Хайло А.М.
ІДЕНТИФІКАЦІЯ АНТРОПОНІМІВ У ПОВНОТЕКСТОВИХ ДОКУМЕНТАХ…….. 30

Усов М.В.
ТИПИ СЛОВНИКІВ: ТЕХНІЧНА ГАЛУЗЬ (СИСТЕМНИЙ ОГЛЯД)…… 32

Варешнюк І.В.
ОБЗОР МЕТОДОВ ФИЛЬТРАЦИИ СПАМА ЭЛЕКТРОННОЙ КОРРЕСПОНДЕНЦИИ…….. 34

Охріменко Ю.М.
ПІДБІР ТЕСТОВИХ ЗАВДАНЬ З ГРАМАТИКИ АНГЛІЙСЬКОЇ МОВИ ДЛЯ СТУДЕНТІВ ПЕРШОГО КУРСУ ФІЛОЛОГІЧНИХ ФАКУЛЬТЕТІВ (НА МАТЕРІАЛІ ГРАМАТИЧНИХ ТЕМ: THE PERFECT TENSES)……...36

Король О.И.
МОДЕЛЬ ИНТЕЛЛЕКТУАЛЬНОЙ ОБРАБОТКИ ПАТЕНТНО-КОНЪЮНКТУРНОЙ ИНФОРМАЦИИ…… 38

Чалова С.Ю.
РОЗРОБКА ТЕСТОВИХ ЗАВДАНЬ З ПРЕДМЕТУ «ЛІНГВОКРАЇНОЗНАВСТВО» (ТЕМА: “HOLIDAYS, SYMBOLS AND TRADITIONS”)……..40

Лой А.А.
ИСПОЛЬЗОВАНИЕ НАИВНОГО МЕТОДА БАЙЕСА ДЛЯ КЛАССИФИКАЦИИ КОЛЛЕКЦИИ ТЕКСТОВ……42

Shevchenko Anna
LINGUISTIC AND CULTUROLOGICAL ASPECTS OF ENGLISH-LANGUAGE COMMUNICATION IN THE FIELD OF EDUCATION……44

Петрасова С.В.
ВЫЯВЛЕНИЕ СЕМАНТИЧЕСКИХ ЭКВИВАЛЕНТОВ ПРИ АВТОМАТИЧЕСКОЙ ОБРАБОТКЕ ТЕКСТОВ…… 46

Коняева К.Г.
СЕМАНТИЧЕСКИЙ АНАЛИЗ КОНЦЕПТОВ «ЛЮБОВЬ» В РУССКОМ И «LOVE» В АНГЛИЙСКОМ ЯЗЫКАХ…… 48

Иванющенко В.С.
ИСПОЛЬЗОВАНИЕ ОНТОЛОГИЙ ДЛЯ ПРЕДСТАВЛЕНИЯ ПРЕДМЕТНОЙ ОБЛАСТИ…….. 50

Ляхвацкая О.Н.
МЕТОДЫ ОБРАБОТКИ СТРУКТУРИРОВАННЫХ ДОКУМЕНТОВ…… 52

Цыбизова Ю.С.
ЛЕКСИКОГРАФИЧЕСКИЙ АСПЕКТ
ОПИСАНИЯ ЛЕКСИЧЕСКИХ ПАРАЛЛЕЛЕЙ…… 54

Прогляда Я.В.
ПРОЕКТУВАННЯ МУЛЬТІМЕДІЙНОЇ НАВЧАЛЬНОЇ СИСТЕМИ ДЛЯ ВИВЧЕННЯ АНГЛІЙСЬКОЇ МОВИ…… 56

Луда С.Э.
ПРОБЛЕМЫ И ОСОБЕННОСТИ ПОСТРОЕНИЯ ИНВЕРТИРОВАННОГО ИНДЕКСА КОЛЛЕКЦИИ ДОКУМЕНТОВ……58

Поморцева Е.Е.
СПЕЦИАЛИЗИРОВАННЫЙ КУРС КАК СРЕДСТВО
ФОРМИРОВАНИЯ КОМПЕТЕНТНОСТИ…… 60

Скапа Л.В.
ОСОБЛИВОСТІ ФУНКЦІОНУВАННЯ НАЙЧАСТОТНІШОЇ ЛЕКСИКИ В АНГЛОМОВНОМУ АКАДЕМІЧНОМУ СПІЛКУВАННІ…… 63

Кулєшова Т.І.
ЗАСТОСУВАННЯ МУЛЬТИМЕДІЙНОГО ОБЛАДНАННЯ НА ЗАНЯТТЯХ З ГРАМАТИКИ АНГЛІЙСЬКОЇ МОВИ…… 65

Данилевич С.Б.
КОМПЬЮТЕРНАЯ ИГРА «ВЕЛИКИЕ ЛЕКСИКОГРАФЫ»……... 67

Періжняк М.М.
ОСНОВНІ ПРОБЛЕМИ ВИКОРИСТАННЯ МАШИННОГО ПЕРЕКЛАДУ ДЛЯ ПЕРЕКЛАДУ ХУДОЖНІХ ТВОРІВ……... 69

Терещенко В.И.
ИСПОЛЬЗОВАНИЕ ОНТОЛОГИЙ ДЛЯ
РЕШЕНИЯ ЗАДАЧ OPINION MINING……... 70

Васильева Ю.В.
РАЗРАБОТКА ЭЛЕКТРОННОГО ПЕРЕВОДЧИКА НА ПЛАТФОРМЕ МАШИННОГО ПЕРЕВОДА APERTIUM……72

Архипенко Л.М.
ДО ПРОБЛЕМИ СТАНОВЛЕННЯ СИСТЕМИ УКРАЇНСЬКОЇ ТЕРМІНОЛОГІЇ ІНТЕЛЕКТУАЛЬНОЇ ВЛАСНОСТІ……74

ПРИКЛАДНОЙ ЛИНГВИСТИКЕ В НТУ «ХПИ» И КАФЕДРЕ ИНТЕЛЛЕКТУАЛЬНЫХ КОМПЬЮТЕРНЫХ СИСТЕМ – ПЯТЬ ЛЕТ!

Шаронова Н.В.
Национальный технический университет "ХПИ",
г. Харьков, ул. Фрунзе, 21, тел. 057 707 64 60,
e-mail: *nvsharonova@mail.ru*

Конференция этого года посвящена пятилетнему юбилею кафедры интеллектуальных компьютерных систем (ИКС). Приказ о создании кафедры был подписан Ректором НТУ «ХПИ» 12 февраля 2007 года. Кафедра с таким названием одна в Украине, она является выпускающей и готовит бакалавров и специалистов по специальности «Прикладная лингвистика». История возникновения кафедры ИКС связано с кафедрами АСУ и Делового иностранного языка и перевода. Модель специальности с самого начала была разработана таким образом, чтобы выпускник овладел профессией, связанной с лингвистическими компьютерными технологиями, наряду с хорошим знанием двух европейских языков (английского и немецкого).

Возникает вопрос: актуальна ли для нашей страны такая специальность и такое направление кафедры? Во всем мире развитие современных информационно-компьютерных технологий требует все большего применения естественного языка не только в качестве универсального информационного интерфейса, но и способа сделать компьютерные системы более интеллектуальными. Эффективность систем автоматизации и информатизации зависит от уровня их интеллектуализации, а собственно языково-мыслительная деятельность человека является сосредоточением его интеллекта и объектом моделирования при построении искусственных интеллектуальных систем.

Традиционные методы в области искусственного интеллекта в определенной мере исчерпали себя, и назрела настоятельная потребность в разработке принципиально новых подходов к созданию интеллектуальных систем. Изучение естественного языка и исследование человеческого интеллекта всегда были самыми актуальными направлениями разработок в области информатики и искусственного интеллекта. В мире в последнее время чрезвычайно активизировались исследования взаимосвязей между языком и мышлением, разработка таких технологических средств, которые адекватно могли бы отображать интеллектуальные возможности человека. Лингвистические технологии признаны в США одним из четырех основных направлений перехода к информационному обществу (тремя другими являются: глобальный доступ, обучающие технологии и цифровые библиотеки), что подтверждает важность подобных исследований. В мире расширяются также исследования, связанные с изучением и моделированием работы мозга, с изучением и моделированием его отдельных функций, в частности, языковых и функций понимания.

Собственно на пересечении информационных систем и языковых исследований лежит область лингвистических технологий, называемая прикладной лингвистикой. Уходя от терминологических споров, которые до сих пор происходят вокруг понятия «прикладная лингвистика», автор, как человек, причастный к открытию специальности «Прикладная лингвистика» в НТУ «ХПИ» пять лет назад, придерживается концептуальной модели специалиста по лингвистическим информационным технологиям с компетенцией в двух-трех иностранных языках. Именно такая модель прикладного лингвиста и отрабатывается на кафедре интеллектуальных компьютерных систем НТУ «ХПИ», где, открывая специальность, старались учесть опыт ведущих высших учебных заведений, как отечественных, так и зарубежных.

Если наша страна хочет быть включенной в европейское информационное пространство, нужно уделять особенное внимание тем сферам науки и образования, которые связаны с лингвистическими технологиями, и внедрять в учебные планы подготовки специалистов по прикладной лингвистике в Украине учебные дисциплины, направленные на получение знаний и умений для возможности выполнения таких исследований. Подготовка будущих прикладных лингвистов должна обязательно сопровождаться углубленным изучением иностранных языков и языковедческих дисциплин (как фундаментальных теоретических, так и новых направлений лингвистических исследований, таких как коммуникативная лингвистика, психолингвистика, когнитивная наука и т.п.), изучением необходимой (дискретной) математики, ряда современных компьютерных дисциплин. Такая стратегия даст возможность готовить специалистов для решения целого ряда актуальных задач в областях, предусматривающих описание и моделирование фонетической, грамматической, семантической и синтаксической структур различного типа текстов, создания словарей, разработки новых методик преподавания иностранного языка и информатики. Такого специалиста требует рынок труда Украины. У такого выпускника есть будущее, в том числе и перспектива заниматься научными исследованиями, защищать кандидатские и докторские диссертации.

За прошедшие пять лет удалось сделать немало. У нас состоялся первый выпуск бакалавров, готовится к защите первый выпуск специалистов. Мы успешно прошли лицензирование специалистов и проходим аккредитацию. Сотрудники кафедры не отступают от своей концепции подготовки специалистов по прикладной лингвистике, стараясь сохранить качество обучения на должном высоком уровне. Наши преподаватели активно трудятся над разработкой учебных планов и пособий, занимаются научной деятельностью, связанной с развитием научного направления «прикладная лингвистика» в Украине. Мы очень надеемся, что наши выпускники будут востребованы нашей страной.

ВИКОРИСТАННЯ ОНТОЛОГІЙ ДЛЯ СЕМАНТИЧНОГО ПОШУКУ ДОКУМЕНТІВ

Дорошенко А. Ю.

Національний технічний університет
"Харківський політехнічний інститут",
м. Харків, вул. Пушкінська, 79/2, тел. 0932468666,
e-mail: marykate90@mail.ru

Онтологія сприяє встановленню коректних зв'язків між значеннями елементів області тексту, тим самим створюючи умови для їх спільного використання. Онтології можна застосовувати в якості будівельних блоків компонентів баз знань, схеми об'єктів в об'єктно-орієнтованих системах, концептуальної схеми баз даних, структурованого глосарію взаємодіючих частин, словника для зв'язку між агентами, визначення класів для програмних систем.

Основні завдання, які успішно вирішуються на базі онтологій, включають надання знань для виведення інформації, що відповідає запиту користувача, фільтрацію і класифікацію інформації, індексування зібраної інформації, організацію загальної термінології, якою можуть користуватися для комунікації програмні агенти і користувачі. Завдання семантичного пошуку в електронній бібліотеці є спрощеним аналогом пошуку інформації в Інтернет, таким чином передбачається, що пошук здійснюватиметься по запиту користувача природною мовою в аналогічному рядку пошуку.

Мета дослідження – запропонувати методику семантичного пошуку, дозволяючи відбирати близькі за загальним контекстом документи, навіть якщо вони належать до різних предметних областей, тим самим можна збирати і узагальнювати знання, розосереджені в різних областях з допомогою семантичного пошуку на основі онтологій.

Пропонуємо методику дослідження, яка полягає у наступних етапах.

1. Методологія побудови онтологій.

Позначення цілей і сфери застосування створюваної онтології. Для цього необхідно визначити, для чого створюється онтологія, і як вона надалі використовуватиметься. Побудова онтології включає в себе фіксацію знань про предметну область, яка полягає у визначенні основних понять і їх взаємин у вибраній предметній області; створенні точних несуперечливих визначень для кожного основного поняття і відношення; визначенні термінів, які пов'язані з цими термінами і стосунками; остаточному узгодженні усіх вище названих етапів. Кодування, яке має на увазі розподіл сукупності основних термінів, використовуваних в онтології, на окремі класи понять; вибір або розробку спеціальної мови для представлення онтології; безпосередньо завдання фіксованої концептуалізації на вибраній мові представлення знань. Нині існує безліч проектів, в основу яких покладені онтології.

2. Семантичний пошук.

«Семантичний пошук – вид автоматизованого повнотекстового інформаційного пошуку з урахуванням смислового змісту слів і словосполучень запиту користувача і пропозицій текстів проіндексованих інформаційних ресурсів» [1].

Як і у будь-якій системі перекладу, виникає завдання зняття омонімії, що правильно відображає зміст текстового висловлювання.

Основна ідея семантичної мережі при вирішенні цієї проблеми полягає в тому, щоб зробити для машини доступну семантику анотацій. Це має бути досягнуто шляхом використання онтологій – концептуальних схем, для надання формального значення термінів, використовуваних в анотаціях, перетворюючи їх на семантичні анотації.

3. Діаграма потоків даних при пошуку. Користувач вводить запит, який підлягає лінгвістичному аналізу, розширюється за рахунок використання синонімів, потім перетворюється в ключові слова і прямує до пошукової машини. Пошукова машина повертає знайдені документи, вони також піддаються лінгвістичному розбору і формує семантичні образи документів. Образи документів порівнюються з образами запиту, робиться висновок про релевантність кожного з документів і результати аналізу надаються користувачеві.

Висновок. Запропонована методика семантичного пошуку дозволяє відбирати близькі за загальним контекстом документи, навіть якщо вони належать до різних предметних областей. Тим самим можна збирати і узагальнювати знання, розосереджені в різних областях за допомогою семантичного пошуку на основі онтологій.

Список літератури

1. *Udo Han*, «Системи семантичного пошуку», 2009. – С. 19-25.

2. *Ланин В.В.* Интеллектуальное управление документами как основа технологии создания адаптируемых информационных систем // Труды международной научно-технической конференций «Интеллектуальные системы» (AIS'07). Т. 2 / М.: Физматлит, 2007. С. 334-339.

3. *Рябова Н., Козополянська Г., Дяденко К.* Застосування онтологічної семантики у зіставленні документів // Матеріали ІІІ Міжн. конф. молодих вчених „Комп'ютерні науки та інженерія" CSE-2009, 14-16 травня, 2009, Львів. – Львів: НУ „Львівська політехніка", 2009. – С. 107-108.

4. *Рябова Н.В.* Методы и модели интеллектуальной обработки текстов в задачах онтологического инжиниринга // Математическое и программное обеспечение интеллектуальных систем (MPZIS-2008). Тез. докл. / YI междунар. науч.-практ. конф., Днепропетровск, 2008. – с.269-270.

НЕЙРОМЕРЕЖНА СИСТЕМА ВАЛІДАЦІЇ ПРОЕКТІВ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ

Говорущенко Т.О.
Хмельницький національний університет,
м. Хмельницький, вул. Інститутська, 11
тел. (095)11-22-544, tat_yana@ukr.net

Вступ. Валідація перевіряє відповідність вимог, проектних рішень, коду програми та результатів функціонування програми потребам користувачів та замовників програмного забезпечення (ПЗ). Валідація дає відповідь на питання: "Чи робимо ми вірний продукт?" [1]. При огляді методів валідації ПЗ слід згадати методологію Safety Case [2]. Узагальнена модель методології Safety Case представлена на рис.1.

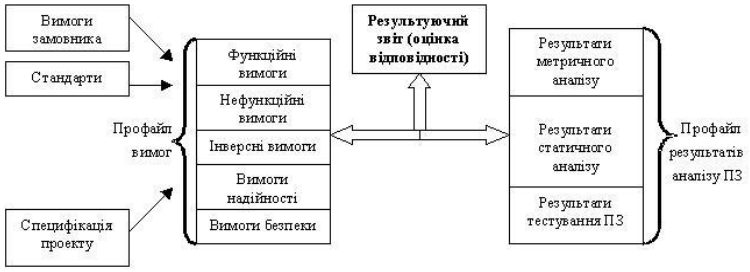


Рис. 1. Узагальнена модель методології Safety Case

Валідації варто піддавати проект ПЗ з метою зниження вартості усунення дефектів у ПЗ. На етапі проектування можливо одержати частину метричної інформації, на основі опрацювання якої потрібно формувати висновки про складність та якість проекту, а також прогноз складності та якості розроблюваного ПЗ. У [3, 4] виділено метрики з точними та прогнозованими значеннями на етапі проектування, а також описано алгоритми їх визначення та діапазони значень.

Нейромережна система валідації проектів ПЗ. Для оцінювання результатів проектування і прогнозування характеристик ПЗ розроблено нейромережну систему валідації проектів ПЗ (НСВП). На вхід НСВП подаються кількісні значення метрик етапу проектування з точними та прогнозованими значеннями, а результатом роботи є висновки про складність та якість проекту та розроблюваного ПЗ. Структурна схема НСВП представлена на рис.2.

НСВП складається з наступних компонентів: 1)діалоговий компонент; 2)блок збирання-передачі даних; 3)база знань; 4)модуль формування вхідних векторів для штучної нейронної мережі (ШНМ); 5)штучна нейронна мережа; 6)модуль опрацювання результатів ШНМ.

Діалоговий компонент візуалізує роботу блоку збирання-передачі даних, відображає роботу системи та видає користувачу повідомлення в зрозумілій для нього формі. *Блок збирання-передачі даних* зчитує інформацію користувача

щодо кількісних значень точних та прогнозованих метрик етапу проектування ПЗ, зберігає одержану інформацію в базі знань та передає її у модуль формування вхідних векторів ШНМ.

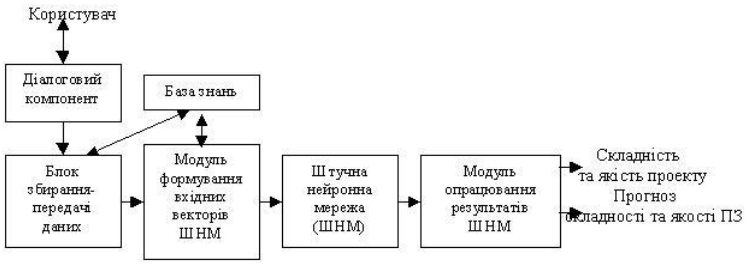


Рис. 2. Структурна схема нейромережної системи валідації проектів ПЗ

База знань містить кількісні значення точних та прогнозованих метрик етапу проектування ПЗ, вхідні вектори ШНМ та правила опрацювання результатів роботи ШНМ. *Модуль формування вхідних векторів ШНМ* готує значення метрик з бази знань до подачі на входи ШНМ. *Штучна нейронна мережа* здійснює апрокисмацію метрик ПЗ етапу проектування та надає кількісну оцінку складності та якості проекту та значення прогнозу характеристик складності та якості розроблюваного ПЗ на основі опрацювання кількісних значень метрик з точними та прогнозованими значеннями на етапі проектування. Архітектура, реалізація, навчання та дослідження ШНМ описані у [3]. На основі 4-х одержаних результатів *модуль опрацювання результатів роботи ШНМ* робить висновки про якість і складність проекту та очікувану якість і складність розроблюваного програмного забезпечення.

Висновки. Одержані оцінки результатів проектування дають дані замовнику для вибору проекту необхідного ПЗ та дозволяють порівняти між собою різні версії проекту, тобто дають змогу прийняти мотивоване та обгрунтоване рішення щодо вибору проекту та його реалізації на основі не лише вартісних та часових характеристик, але й з врахуванням характеристик складності та якості проекту і розроблюваного програмного забезпечення.

Список літератури

1. *А. Карпов.* Верификация и валидация // http://software.intel.com/ru-ru/blogs/2010/02/05/2003055

2. *Górski J.* Trust Case – a case for trustworthiness of IT infrastructures / J. Górski // Cyberspace Security and Defense: Research Issues, 2005.

3. *Поморова О.В., Говорущенко Т.О., Онищук О.С.* Оцінювання результатів проектування та прогнозування характеристик якості програмного забезпечення // Вісник Хмельницького національного університету - Хмельницький: ХНУ, 2011 - №2, с.168-178.

4. *Поморова О.В., Говорущенко Т.О.* Інтелектуальний метод оцінювання результатів проектування та прогнозування характеристик якості програмного забезпечення // Радіоелектронні і комп'ютерні системи – Харків: НАУ "ХАІ", 2010 – № 6, С. 211-218.

GENERALIZING FRAMEWORK FOR ONTOLOGY LEARNING

Orobinska O.O.

Laboratoire ERIC de l'université Lyon-2 Lumière
5 avenue Pierre Mendès-France 69676 Bron cedex,
tel. : +33 (0)4 78 77 23 76 e-mail: *Olena.Orobinska@univ-lyon2.fr*

The techniques of ontology engineering have got new impact during the last decade with the occurrence of the different kinds of digital resources (such as MRDs, lexicons, thesaurus etc.). This has permitted to apply widely linguistic methods towards the tasks of ontology building and has generated new domain of ontology engineering such as ontology learning.

With in reference to M. Bergman [1] in 2011 among the fastest growing categories of "sweet tools"[1] have been all ones related to ontologies with Java as the dominant language of these tools.

But for all the successes in ontology engineering there is neither universal procedure generally accepted nor a fortiori any consistent suites of tools allowing to conceive in a traceable, explicit way an domain ontology from group of informational resources relevant to this domain [2].

We propose a general framework for ontology learning based on the related researches [3] and our point of view in ontology learning. According our approach the different components that are involved in ontology learning process is integrated into the framework to generate an ontology.

This framework consists of three backbone stages: preliminary information extraction, ontology discovery, and ontology organization.

Preliminary information extraction. A variety of data can be exploited in ontology learning, including raw-text documents (e.g., scientific articles, Web pages, book chapters, newspaper), (semi-)structural data (e.g., Web site structure lexicons), and usage data (e.g., the log of user navigation and search queries). Information extraction pre-processes and recognizes information in a variety of forms and converts them into the forms that can be used for ontology discovery. In particular, text documents are processed via content analysis by employing a variety of natural language processing techniques, ranging from tokenization, to part-of-speech tagging, phrase structure and/or grammatical function parsing, semantic and discourse analyses.

Ontology componrnts discovery. Supervised and unsupervised learning algorithms have been applied to discover the concepts and the relations from the extracted information. Approaches to relation learning vary in terms of the scope of co-occurrence, the metrics for the significance of co-occurrence, the criteria for selecting candidate concepts, and the thresholds for extracting potential relations. Some learning approaches require the assistance of domain-specific resources such as thesauri.

[1] semantic Web and -related tools by the AI3

Ontology implementation. Given the large number of possible ontological concepts and relations extracted from the learning process, an issue arises as to how to improve the usability of the discovered knowledge. Ontology implementation seeks to achieve the above goal via the following steps:

• clustering synonymous terms and their relations;

• discovering local centers of concepts. A concept cloud is a group of concepts that are closely related among themselves but marginally related to concepts outside the group. If connect any two concepts *i* and *j* that are directly related within a cloud via path p_{ij}, a network will be constructed. The length of *p* is the total number of concepts on *p* minus 1. Cloud center C_c can be found by looking for the concept(s) that has the shortest total path connecting to the rest of the concepts on the same cloud (the total is M), as shown in (1):

$$C_C = \min \sum_{i=1, i \neq c}^{M} p_{ci} \qquad (1);$$

• deriving inverse relations. We can derive inverse relationships from the ones that have been discovered;

• building higher-level ontology. The local centers discovered in the previous step can serve as the top-level ontology of the target domain. The process for identifying local centers in a concept cloud is repeated to find concepts for lower-level ontology by selecting concepts that have the next shortest length of path. This process continues to the desired level where concepts constitute the leaf nodes of an ontology hierarchy. It is possible to construct such a hierarchy for each relation separately.

It is desirable to automate all the components in the framework for ontology learning by developing techniques.

References

1. *Bergman M.* AI3's Inaugural State of Tooling for Semantic Technologies/ http://www.mkbergman.com/991/the-state-of-tooling-for-semantic-technologies.

2. *Charlet, J.,* at al. Apport des outils de TAL a la construction d'ontologies : propositions au sein de la plateforme DaFOE/ Charlet, J., at al dans TALN 2009, Senlis, 24–26 juin 2009

3. *De Nicola, A.,* & all. A software engineering approach to ontology building/ De Nicola, *A.*, et al In Information Systems 34 (2009) 258–275.

МОДЕЛИРОВАНИЕ ОСНОВНЫХ ПОЛОЖЕНИЙ АППРОКСИМИРОВАННОЙ СТРУКТУРЫ СОДЕРЖАНИЯ ХУДОЖЕСТВЕННОГО ПРОИЗВЕДЕНИЯ

Неручок Ю. Ю.
Национальный технический университет "Харьковский политехнический институт",
г. Харьков, ул. Пушкинская, 79/1, тел. 0952505464,
e-mail: *good-badgirl@mail.ru*

На современном этапе развития лингвистики, в эпоху всеобщей компьютеризации и глобальной информатизации наш интерес к опыту точных наук и использование этого опыта в разработке авторской технологии исследования текста можно считать отвечающим запросам времени.

Выявление семантически значимой информации происходит во многом благодаря использованию средств компьютерной графики. Высокая эффективность графического представления информации подтверждена многочисленными исследованиями наглядно-образного и визуального мышления на основе когнитивного подхода. Фундаментальная идея этого подхода заключается в том, что мышление есть не что иное, как манипулирование внутренними (мысленными) репрезентациями структурированных определенным образом знаний – фреймов, сетей, планов, сценариев. Исходя из этого, разработаны модели представления знаний на основе семантических сетей, продукций, логики предикатов и нечетких знаний.

Цель исследования – разработка и применение метода анализа художественного текста как системы таких алгоритмизируемых процедур, которые способны выявлять лингвистически значимую информацию, моделировать процесс перехода от лексико-грамматического уровня текста к структуре его содержания, визуализировать статику и динамику последнего.

Объектом анализа является содержание художественного текста как ментальное образование – семантика текста в её статике и становлении.

Материалом для исследования послужил рассказ А.П. Чехова «Дама с собачкой», хорошо структурированный небольшой по объему текст (его длина составляет 317 предложений) как обозримое целое для анализа, но с достаточным материалом для классификации.

Предлагаемый метод исследования семантики текста предполагает пошаговое решение следующих этапов:

1. Первым этапом анализа является выделение ключевых слов, опирающееся на анализ употребительности слов в тексте.
2. Вторым, к сожалению, пока не формализованным этапом анализа является установление кореферентного тождества слов и словосочетаний с целью получения полной информации о номинировании персонажей (действующих лиц, участников ситуации).

3. Формализация семантического пространства может осуществляться через анализ синтаксических связей членов предложения.

4. Количество минимально необходимых для существенного представления содержания художественного текста в его статике и динамике предикативных типов равно 11: акциональный, перцептивный, экзистенциальный, сущностный, квалификативно-дескриптивный, социативный, посессивный, коммуникативно-контактный, предикаты речи, движения и состояния.

5. Динамика отношений между персонажами (количество и качество их отношений) может измеряться через количество и качество их синтаксических связей, опосредованных глаголами указанных семантических типов.

6. Семантическая сеть текста может использоваться для выделения значимых частей текста как условие моделирования динамики его содержания.

Теоретическая значимость работы состоит в попытке выйти за рамки анализа художественного текста, который является на данном этапе исследований пока традиционным и собственно лингвистическим, и разработать алгоритмический метод использования лингвистической информации для выявления и наглядного представления содержания текста.

Практическая ценность предложенного алгоритма заключается в возможности как его самостоятельного использования, так и в составе других алгоритмических процедур для решения ряда практических задач, связанных с компьютерной обработкой текста в рамках литературо- и переводоведения, а также для построения интегрированных технологий для систем искусственного интеллекта.

Список литературы

1. *Новиков А.И.* Семантика текста и его формализация / А.И. Новиков. – М.: Наука, 1983. – 361 с.

2. *Поспелов Д.А.* Моделирование рассуждений /Д.А. Поспелов. – М.: Наука, 1989. – 215 с.

3. *Чехов А.П.* Дама с собачкой / А.П. Чехов. – Электр, текст – 10 с.

4. *Апресян Ю.Д.* Современные методы изучения значений и некоторые проблемы структурной лингвистики / Ю.Д. Апресян / Проблемы структурной лингвистики. – М.: Изд-во АН СССР, 1963. – 102-150 с.

5. *Марчук Ю.Н.* Основы компьютерной лингвистики / Ю.Н. Марчук. – М.: Изд-во МПУ «Народный учитель», 2000. – 226 с.

ЕКСПЕРИМЕНТАЛЬНА ФОНЕТИКА У СВІТЛІ СУЧАСНИХ КОМП'ЮТЕРНИХ ТЕХНОЛОГІЙ

Іщенко О.С.
Інститут української мови НАН України,
Київ, вул. Грушевського, 6, тел. 0 (44) 278–18–85,
e-mail: o.ishenko@gmail.com

Фонетика – це наукова дисципліна, яку присвячено дослідженню звуків та інших явищ усного мовлення. Її називають крос-галузевою ділянкою мовознавства, адже, розв'язуючи лінгвістичні проблеми, фонетисти активно залучають знання з фізики, анатомії, нейрофізіології, психології, кібернетики тощо. У свою чергу, окремим розділом фонетичних знань є експериментальна фонетика, в якій дослідження здійснюють емпіричними методами; серед них провідний – експеримент. Емпіричні методи фонетичного аналізу дають об'єктивну оцінку кількості та якості фонетичних одиниць мови, їх взаємодії в потоці мовлення, виявляють звукову видозміну, варіативність, засвоєння, вплив, дозволяють моделювати мовленнєві явища. Експериментальна фонетика реалізує перевірку теоретичних гіпотез і припущень. Її результати стають об'єктом уваги не лише лінгвістів, а й фахівців із синтезу та розпізнавання усного мовлення, передавання інформації каналами зв'язку, криміналістичної експертизи, лікарів, психологів, педагогів та ін.

Сучасна наукова емпірична база дає широкі можливості для глибоких і потужних досліджень. Новітні програмно-технічні засоби (комп'ютерні технології) забезпечують якість і швидкість пошуку, збору, аналізу, передавання, зберігання, накопичення та відтворення інформації. Тому комп'ютеризація у сфері експериментальної фонетики сприяє високій надійності та оперативності дослідницької роботи й розширює її межі.

Експериментально-фонетичні дослідження вимагають цілий комплекс просторових і технічних умов – лабораторію. Просторові умови пов'язані передусім зі спеціальним акустично підготовленим приміщенням для запису мовлення – студією. Приміщення аудіостудії повинне характеризуватися звукоізоляцією та звукопоглинанням. Технічні умови передбачають забезпечення запису звукового сигналу, його прослуховування, інструментального опрацювання й аналізу. Лабораторій експериментальної фонетики (ЛЕФ) в Україні, на жаль, обмаль.

Для запису та вивчення мовлення необхідна звукова апаратура. На сьогодні цілком достатньо мікрофона та персонального комп'ютера. ПК повинен мати звуковий адаптер і драйвер звукового адаптера, далі – необхідний комплекс програмного забезпечення. Натомість раніше апаратура мала широкий склад: спектрографи, смугові аналізатори, катодні й шлейфні осцилографи, інтонографи, магнітофони тощо.

Новітні комп'ютерні програми дають змогу записувати мовлення, прослуховувати, редагувати та аналізувати його. Серед професійного

програмного забезпечення назвемо: Aneto (Technical University of Catalonia), Audition (Adobe Incor.), Praat (University of Amsterdam), Sigview (SignalLab Co.), Sonic Visualiser (University of London), Sound Forge (Sony Corp.), Speech Analyzer (SIL Org.), WASP/SFS (University College London), Wavelab (Steinberg Co. Ltd), Wavesurfer (KTH of Stockholm), WinCecil (SIL Org.) тощо, до більшості з яких є безкоштовний чи умовно безкоштовний доступ в Інтернеті[2]. Такі програми дозволяють побачити акустичну хвилю, вивчати її протяжність, амплітуду коливань, спектральні компоненти – основну частоту, гармоніки.

Предметом фонетичних досліджень є не лише звукове мовлення, а й власне артикуляція. Процес мовлення спостерігають за допомогою інших інструментальних засобів, зокрема фотографування, фільмування, рентгенографування, палатографування (фіксація контактів язика з твердим піднебінням), міографування (фіксація біопотенціалів м'язів артикуляційних органів), ларингографування (фіксація вібрацій голосових зв'язок і рухів гортані), кімографування (фіксація сили тиску повітряного струменя), назометрографування (фіксація акустичної енергії ротової і носової порожнин окремо) тощо. Безперечним здобутком української артикуляційної фонетики є прийом тензопалатоосцилографування, винайдений Л. Скалозуб і В. Лебедєвим. Щоправда, новітні дослідження звукотворення провадять здебільшого за допомогою електромагнітного артикулографа, що реєструє рухи артикуляторів у тривимірній площині (3d-технологія), магнітно-резонансного томографа та електропалатографа.

Попри широкі можливості для вивчення звуко-мовленнєвих явищ фонетика української мови має багато лакун та нерозв’язаних актуальних проблем. Сподіваємося, що представлений огляд можливостей експериментального аналізу сприятиме ангажуванню нових дослідників до цікавої наукової царини – фонетики.

[2] Ознайомитись із програмним забезпеченням, створеним для аналізу та редагування записів усного мовлення, можна на таких веб-ресурсах: <http://www.speechandhearing.net/laboratory/tools.php> або <http://www.phonetica.wordpress.com/links>.

СТВОРЕННЯ ТЕСТІВ З ГРАМАТИКИ АНГЛІЙСЬКОЇ МОВИ ЗА ГРАМАТИЧНИМИ ТЕМАМИ«REPERTED SPEECH» ТА «SEQUENCES OF TENSES»

Михайлов М.С.
Національний аерокосмічний університет ім. М.Є. Жуковського «Харківський авіаційний інститут»
г. Харків, вул. Чкалова, 17, тел. 778-40-09,
e-mail: 777lancer777@ukr.net

Сучасний освітній простір створює певні пріоритети у навчанні. Завдяки реформам у сфері вищої освіти практично всі ВНЗ зорієнтовані на використання тестів для остаточного виміру показника знань студентів. Практика застосування тестів напрацьовувалася більш ніж століття. Під тестом розуміють модернізовану систему вимірювання знань студентів, що, в свою чергу, складається з системи тестових завдань, структури проведення та організації, обробки результатів та їх аналізу [1, с. 311].

Дослідження показують, що проведення тесту вимагає ретельного підбору завдань. Тест як система має свої функції та класифікацію як завдань закритого і відкритого типів, які мають певні підтипи[4, с. 23].

Із забезпеченням правильної організації тестування сприяє розвитку пам’яті, мислення та мови студентів, систематизує їхні знання, своєчасно викриває прорахунки навчального процесу та служить їх запобіганню, тому контроль та оцінка знань, умінь та навичок студентів є важливим елементом навчально-виховного процесу[2, с. 134].

Тестовий контроль спрощує перевірку робіт викладачем і дає змогу організувати рубіжний та підсумковий контроль, активізувати діяльність студентів шляхом охоплення контролем більшої їх кількості, перевірити знання великого за об’ємом матеріалу за невеликий проміжок часу[3, с. 56].

Об’єктом нашого дослідження є загальні положення тестології, а *предметом* – вивчення головних етапів процесу створення тестів для перевірки знань студентів з граматики англійської мови за певними граматичними темами «ReportedSpeech»та «SequencesofTenses», їхніх функцій та стилів.

Метою є розробка тестових завдань для перевірки рівня знань студентів філологічних спеціальностей граматики англійської мови за навчальними темами «ReportedSpeech»та «SequencesofTenses» та їх подальше використання з метою максимально чіткого оцінювання рівня знань студентів.

Матеріал дослідження базується на методичних збірниках, а також основних положеннях теоретичної граматики вітчизняних та іноземних вчених.

Практична частина нашого дослідження полягає у розробці тестових завдань на зазначені вище граматичні теми. Важливими вимогами до створення тестових завдань є наступні:

1. Обов’язковий добір тестових завдань, пов’язаних із планованою для тесту змістовою валідністю.

2. Дотримання тієї умови, що тести успішності мають обмежену галузь застосування, яка, як правило, розповсюджується на окремі стадії навчання (навчальний рік, етапи навчальної програми). Умовою розробки і стандартизації таких тестів є опора на стандартні програми навчання.

Нами опрацьовано значну частину теоретичного матеріалу, що стосувався підбору тестів та правил граматики англійської мови за заданими темами. Це дало можливість відібрали саме той граматичний мінімум (з урахування стандартної програми навчання), котрий обов'язково повинні знати студенти для проходження тестів.

Було розроблено чотири тести з курсу граматики англійської мови за темами «SequencesofTenses» та «ReportedSpeech». Тести переважно відкритої структури та змішаної форми, оскільки ці граматичні теми дуже взаємопов'язані. Наприклад, завдання закритого типу включають завдання з вибором відповіді; відкритого типу – на заповнення пропусків. Кожен з тестів має однакову кількість різних завдань – 50.

Також було створено ключі до тестових завдань.

Список літератури

1. *Аванесов В.С.* Научные проблемы тестового контроля знаний: Монография / В.С. Аванесов. – М.: МИСиС, 1994. – 278с.

2. *Белий Ю.О., РапопортІ. О.* Тесты как инструмент и как объект педагогических исследований. Объективные характеристики, критерии, оценки и измерения педагогических явлений и процессов. – М., 1973. – 420с.

3. *Бонди Е. О.* Языковые тесты и тестирование (на материале английского языка) / Вопросы лингвистики и методики преподавания иностранных языков. Вып. 2. – М., 1972.–306с.

4. *Бризгалова В. Г., Драчева Г. И.* Опыт составления и использования тестов по грамматике английского языка в техническом вузе / Проблема контроля при обучении иностранным языкам в вузе. Вып. 1. – Таганрог, 1972. –245с.

МЕТОДИ І МОДЕЛІ ОБРОБКИ ІНФОРМАЦІЙНИХ ОБ'ЄКТІВ У СУЧАСНИХ БІБЛІОТЕЧНИХ СИСТЕМАХ

Кочуєва З.А.

Національний технічний університет "ХПІ",
м. Харків, вул. Фрунзе, 21, тел. (057)707-63-60,
e-mail: kochueva@mail.ru

Інформація є одним із найбільш значимих ресурсів, які впливають на розвиток суспільства, його культуру, науку. Бібліотека завжди була центром наукового та культурного життя людей. Сьогодні, з бурхливим розвитком інформатизації освіти, роль бібліотек не зменшується, а навпаки збільшується, а бібліотеки шукають нові форми роботи, які дозволяють здійснювати все більший спектр інформаційної та просвітницької діяльності.

З розвитком мережі електронних бібліотек все більшої актуальності набувають задачі переробки інформації, яка міститься у бібліотеках, розташованої як на традиційних носіях у вигляді книг, газет, журналів, так і представленої в електронному вигляді на сучасних носіях інформації. Основою створення автоматизованих інформаційних бібліотечних систем (АІБС) є розробка математичного і лінгвістичного забезпечення, а саме: моделей, алгоритмів і методів, які охоплюють процеси предметизації, анотування та реферування, при цьому проблеми, що розглядаються, можна класифікувати таким чином: ідентифікація, представлення та обробка знань у АІБС; розробка природномовного інтерфейсу АІБС із залученням методів та засобів комп'ютерної лінгвістики; імітація процесу «розуміння» при аналітико-синтетичній обробці документу, для чого необхідним є залучення знань з моделювання інтелектуальних функцій людини.

У межах окресленої проблеми важливими є наукові задачі розробки моделей, методів, алгоритмів та програм, які здійснюють моделювання процесів інтелектуальної обробки інформаційних об'єктів з метою визначення їх основних характеристик для побудови інформаційного, математичного, лінгвістичного і програмного забезпечення АІБС. У вирішенні задачі ідентифікації знань істотний внесок внесли вчені В.М. Глушков, А.К. Жолковський, Ю.М. Марчук, М. Мінський, О.В. Палагін, Д.О. Поспєлов, Р.Ш. Рубашкін, Ч.Дж. Філлмор, Н. Хомський, Р. Шенк, Я.Л. Шрайберг та ін.

На основі розроблених у роботі методів і моделей інтелектуальної обробки інформаційних об'єктів і процесів у сучасній бібліотеці запропоновано інформаційну технологію формалізації процедур систематизації і предметизації повнотекстових документів, анотування й реферування документів, сформульовано критерії ефективності інформаційного пошуку в електронній бібліотеці, запропоновано розв'язання задачі книгозабезпеченості, розроблено систему ретроспективної конверсії каталогів. Математичні результати роботи можуть бути використані в системах автоматичної обробки природної мови,

при розробці різних інформаційно-пошукових, експертних, аналітичних засобів інформаційних систем широкого призначення [1, 2].

Метою дослідження є підвищення ефективності роботи сучасної бібліотеки за рахунок використання інтелектуальних методів і моделей обробки інформаційних об’єктів. Відповідно до зазначеної мети поставлено та розв’язано такі задачі:

1) виконано аналіз методів і моделей представлення і обробки знань, ідентифікації інформаційних об’єктів при автоматизації інформаційно-бібліотечних систем і сформульовано основні вимоги до розробки їхнього лінгвістичного забезпечення;

2) розроблено математичні та лінгвістичні засоби для розв'язання задач обробки текстових документів на основі моделювання лінгвістичної діяльності людини і інтелектуального аналізу даних методом компараторної ідентифікації;

3) розроблено модель знання-орієнтованого аналізу у задачах анотування та реферування повнотекстових документів;

4) удосконалено модель процесу систематизації і предметизації повнотекстових документів у бібліотечних системах;

5) виконано практичну реалізацію запропонованих методів і математичних моделей, впроваджено результати дисертаційної роботи у практику створення інформаційних бібліотечних систем.

Об’єктом дослідження є інформаційні об’єкти і процеси в автоматизованих інформаційних бібліотечних системах.

Предметом дослідження є методи і моделі інтелектуальної обробки інформаційних об’єктів і процесів.

Методи дослідження засновані на комплексному використанні теорії інтелекту, алгебри скінчених предикатів та предикатних операцій, методу компараторної ідентифікації, методів штучного інтелекту для розробки алгебро-логічних моделей обробки інформаційних об’єктів, які представлено текстовою формою. Алгебра предикатів та предикатних операцій використовується для формалізації знань, опису природно-мовних відношень та моделювання процесів анотування та реферування повнотекстових документів. Метод компараторної ідентифікації використовується для опису інтелектуальних функцій людини при аналітико-синтетичній обробці документів.

Список літератури

1. *Алисейко, З. А.* Программная поддержка автоматического реферирования текста [Текст] / З. А. Алисейко, А. Б. Кованев, Н. В. Шаронова // Вестн. Херсон. нац. техн. ун-та. – Херсон : ХНТУ, 2005. – № 1(21). – С. 331-334.

2. *Алисейко, З. А.* Моделирование систематизации и предметизации полнотекстовых документов [Текст] // Проблемы информационных технологий / З. А. Алисейко, О. В. Канищева, Н. В. Шаронова // Проблемы информационных технологий. – Херсон : ХНТУ, 2007. – № 1. – С. 140-144.

РЕАЛИЗАЦИЯ ВЫЯВЛЕНИЯ И ИСПРАВЛЕНИЯ ОРФОГРАФИЧЕСКИХ ОШИБОК В ТЕКСТЕ

Агеев И.В.
Национальный технический университет "Харьковский политехнический институт",
г. Харьков, ул. Пушкинская, 79/2, тел. 707–63–60,
e-mail: jason4eg@gmail.com

Реальностью сегодняшнего дня стали электронные издания, число которых постоянно увеличивается. Библиотеки, не имеющие в фонде тех или иных электронных изданий и предоставляющие к ним доступ через Интернет, уже с полным на это основанием включают их библиографические описания в свои каталоги и предоставляют их пользователям.

При создании различных по назначению баз данных производится ввод текстовой информации, осуществляемый двумя способами – набором вручную или сканированием. В обоих случаях возможны орфографические ошибки.

Современные текстовые редакторы (например, MS Word), как правило, проводят автоматизированную проверку орфографических ошибок. Это требует вмешательства пользователя. Автоматическая коррекция орфографических ошибок может быть более эффективным средством минимизации опечаток и их исправлений при создании текстовых файлов.

В данной работе была поставлена задача, рассмотреть существующие методы автоматической коррекции орфографических ошибок.

Орфографически ошибочным словом называется буквенная цепочка, полученная некоторым преобразованием из словоформы некоторой лексемы, принадлежащей естественному языку. Под исправлением ошибки в таком слове называется установление исходной словоформы. Исходная словоформа определяется неоднозначно. В данной постановке задача исправления ошибки называется также задачей полного словарного исправления. Результатом попытки исправления ошибки в пределах некоторого класса преобразований может быть также установление невозможности ее исправления, то есть несуществования в словаре словоформы, из которой данная образуется цепочка путем какого-либо преобразования заданного класса [1].

Опечатками называются ошибки, связанные с поверхностным, буквенным представлением слова, то есть с искажениями непосредственно буквенной цепочки, представляющей словоформу [1].

С целью объяснения и исправления ошибок выделяется некоторый класс элементарных искажений. Например, замена одной буквы на другую, перестановка гласной буквы через согласную, сдвиг руки на одну позицию при набивке части слова на клавиатуре. Сложными называются ошибки, являющиеся комбинацией нескольких элементарных. Например, замена двух букв в одном слове. Элементарное искажение называется локальным, если оно по определению затрагивает небольшой отрезок буквенной цепочки, например,

не больше трех букв. Цепочка называется словом с одиночной ошибкой, если она содержит только одно элементарное искажение [1].

Одним из важнейших классов элементарных искажений является класс однобуквенных ошибок, включающий в себя:

1. Замена одного символа на другой (83%)
2. Удаление символа (8%)
3. Добавление лишнего символа (4.5%)
4. Перестановка двух символов (4.5%)

В процессе анализа методов выявления и устранения орфографических ошибок были выделены следующие методы:

- метод n-грамм
- метод спел-чекера

Метод n-грамм основан на предположении, что похожие слова обладают достаточным количеством общих подстрок длины n (n-грамм). При создании индекса для каждого слова составляется список содержащихся в нем n-грамм, который сохраняется в инвертированном виде. При такой организации данных для каждого указателя в инвертированном списке n-грамм нужно в произвольном порядке считывать ключевые слова из словаря, но скорость произвольного доступа во много раз меньше чем последовательного. Еще одна проблема связана с поиском по коротким терминам, когда изменение одной буквы приводит к «непопаданию» слова в выборку. Метод основан на том факте, что каждом языке существует строго ограниченный набор допустимых сочетаний символов.

Метод спел-чекера часто применяется в системах проверки орфографии (т.е. в spell-checker'ax), там, где размер словаря невелик, либо же где скорость работы не является основным критерием. Он основан на сведении задачи о нечетком поиске к задаче о точном поиске. Из исходного запроса строится множество «ошибочных» слов, для каждого из которых затем производится точный поиск в словаре. Алгоритм может быть легко модифицирован для генерации «ошибочных» вариантов по произвольным правилам, и, к тому же, не требует никакой предварительной обработки словаря, и, соответственно, дополнительной памяти. Можно генерировать не всё множество «ошибочных» слов, а только те из них, которые наиболее вероятно могут встретиться в реальной ситуации, например, слова с учетом распространенных орфографических ошибок или ошибок набора.

Список литературы

1. *Гельбух А.Ф.* Эффективно реализуемая модель морфологии флективного естественного языка. /Гельбух А.Ф. – Москва, 1994. – 77 с.

ИЗВЛЕЧЕНИЕ КЛЮЧЕВЫХ ПОНЯТИЙ ИЗ ТЕКСТОВЫХ ДОКУМЕНТОВ ПРИ ГЕНЕРАЦИИ ПРЕСС-ПОРТРЕТА

Гостева Е. С.
Национальный технический университет
«Харьковский политехнический институт»
г. Харьков, ул. Пушкинская, 79/2, тел.707-63-60
e-mail: lisa.gosteva@ gmail.com

Одним из вариантов решения проблемы идентификации фактов в текстах и извлечения их характеристик является использование набора образцов. Образцами могут служить возможные лингвистические варианты фактов, которые можно поместить в интегрированный ресурс знаний (РЗ), объединяющий базу предметных знаний и словарь. Такой подход позволяет представить найденные ключевые понятия, представленные событиями и отношениями, в виде структур, которые в том числе можно хранить в базах данных [1-4].

Лингвистическая составляющая ресурса знаний — словарь. Словарь связан с базой предметных знаний посредством ссылок от дескрипторов к элементам знаний: дескрипторы словаря базовой лексики ссылаются на концепты, а дескрипторы словаря собственных имен — на априори известные экземпляры концептов из базы фактов.

Словарь базовой лексики и словарь собственных имен имеют схожее устройство — это дескрипторные словари (дескриптор представляет множество синонимичных выражений). В отличие от тезауруса, дескрипторы в словаре базовой предметной лексики не связаны друг с другом никакими парадигматическими отношениями (роль последних выполняют отношения между соответствующими элементами базы предметных знаний). В словаре собственных имен словарным входам приписаны довольно общие категории типа «имя лица», «название организации» (такие категориальные метки удобно использовать на этапе извлечения первичных текстовых фактов).

Сложность извлечения фактов с помощью образцов связана с тем, что на практике их нельзя представить в виде простой последовательности слов. Поэтому для идентификации различных уровней компонентов и отношений требуется предварительная обработка естественного языка на разных уровнях: первичная фильтрация документа; лингвистическая обработка: графематика, морфология, синтаксис; выделение простейших семантических структур; собственно извлечение информации и объединение построенных структур или интеграция фактов [2].

На стадии интеграции найденные в документах факты, исследуются и комбинируются. Это выполняется с учетом отношений, которые определяются местоимениями или описанием одинаковых событий. Также на этой стадии делаются выводы из ранее установленных фактов.

Компонент генерации фактов решает две основные задачи: генерацию (порождение) смысла будущей единицы пресс-портрета и лингвистический синтез самого высказывания по порожденному смыслу [3, с. 49-51]. Первый этап генерации фактов включает: определение информации, которая будет формировать пресс-портрет, построение семантической сети (графа), определение последовательности выдаваемой пользователю информации в соответствии с порядком фраз в выходном тексте и определение лексем, которые будут замещать позиции семантической сети конечного текста.

Второй этап процесса генерации конечного текста пресс-портрета связан с построением фраз на естественном языке. Для этого необходимо найти решение для следующих задач: построение синтаксической структуры будущей фразы; определение морфологической информации для входящих в составные части фразы слов; морфологический синтез всех словоформ фразы на естественном языке. Данная информация синтезируется из результатов лексико-синтаксической обработки текста.

Объединение лингвистических и предметных знаний в одном ресурсе, во-первых, облегчает первичное наполнение и последующую поддержку, а во-вторых, дает возможность использовать предметные знания уже на этапе первичной обработки текста правилами извлечения информации. Такой подход позволяет разработать специальный язык запросов к РЗ, при этом правила могут не ограничиваться словарной информацией, а обращаться к семантической сети (или онтологии) и базе фактов для проверки различных условий, требующих навигации по отношениям.

Список литературы

1. *Александровский Д.А., Кормалев Д.А., Куршев Е.П., Сулейманова Е.А, Трофимов Е.В.* Реализация ресурса знаний в системе извлечения информации из текста.// Сборник докладов. — М.: МАКС Пресс, 2007.

2. *Барсегян А. А., Куприянов М С., Степаненко В.В.*Технологии анализа данных: DaraMiningVisualMiningTextMining, OLAP. – СПб.: БХВ-Петербург, 2007. —384 с.

3. *Зубов А.В., Зубова И.И.* Основы искусственного интеллекта для лингвистов: Учеб.пособие. – М.: Университетская книга; Логос, 2007 – 320 с.

4. *Селезнев К.* Обработка текстов на естественном языке.//«Открытые системы», № 12, 2003.

ПОСТРОЕНИЕ И ИСПОЛЬЗОВАНИЕ МАТРИЦЫ ИНЦИДЕНТНОСТИ «ТЕРМИН-ДОКУМЕНТ» В ПОЛНОТЕКСТОВОМ ИНФОРМАЦИОННОМ ПОИСКЕ

Дашкевич Е.С.
Национальный технический университет
«Харьковский политехнический институт»
г.Харьков, ул. Пушкинская 79/2, тел. 0978314534
e-mail: esdashkevich@yandex.ua

В эпоху стремительного развития информационных систем, а также быстро растущих потребностей современного общества, остро встает вопрос о создании специфически-ориентированного программного обеспечения. В условиях современной глобальной информатизации, затрагивающей все сферы человеческой деятельности, каждый пользователь нуждается в мгновенных методах обработки различных видов информации. В первую очередь, это касается текстовых документов. С целью усовершенствования методов обработки информации необходимо рассмотреть существующие методы решения задачи построения матрицы инцидентности «термин-документ» для коллекции документов и разработать программное обеспечения для решения данной задачи.

Информационный поиск – обширная междисциплинарная область науки, стоящей на пересечений большого количества наук. Автоматические системы информационного поиска используются для уменьшения так называемой «информационной нагрузки». Наиболее известными примерами ИП являются методы, используемые в интернете. Большинство современных поисковых систем при организации поиска используют скрытое семантическое индексирование. Появление терминов в документе представляется при помощи матрицы «термин-документ».

Для сортировки текстов в коллекции по релевантности, необходимо определить, соответствует ли данный текст запросу, насколько высоко это соответствие. Для этого может быть использована матрица инцидентности «термин-документ» для коллекции документов.

Матрица инцидентности в широком понимании – одна из форм выражения графа, в которой обозначены святи между инцидентными элементами графа. Столбцы матрицы соответствуют ребрам, строки – вершинам графа. Ненулевое значение в ячейке на пересечении строки и столбца указывает на связь между вершиной и ребром – их инцидентность.

В частном случае (а именно в области информационного поиска) матрица инцидентности представляет собой таблицу, отображающую наличие термина в документе. Принцип заполнения матрицы следующий:

1) столбцами матрицы являются документы из предложенной коллекции документов, строками – термины из данных документов;

2) при наличии термина в тексте, в соответствующую ячейку на пересечении столбца и строки заносится единица; при его отсутствии – ноль.

На основе проведенного анализа был разработан следующий, наиболее рациональный, общий алгоритм решения данной задачи:

- создание массива структур;
- разбиение текстов на лексемы;
- исключение повторений терминов в рамках одного текста;
- анализ повторений терминов во всех текстах;
- заполнение матрицы инцидентности.

Преимущества рассматриваемого общего алгоритма:

- непосредственный доступ к файлу, в котором производится поиск;
- исключение повторений в рамках одного текста предоставляет возможность корректного заполнения матрицы;
- последовательный анализ повторений терминов во всей коллекции документов позволяет быстро заполнить матрицу инцидентности «термин-документ»;
- с помощью построенной таким образом матрицы инцидентности «термин-документ» становится возможным построение инвертированного списка для дальнейшего поиска.

Наряду с рассмотренными преимуществами, данный способ решения имеет и недостаток. Им является высокая сложность обработки большого количества документов коллекции.

После выполнения программы, пользователь получает выходную информацию на экран. Стоит заметить, что пользователь не принимает участие в ходе программы, что исключает появление некоторых ошибок. После получения результата работы программы, предоставляется возможность обрабатывать данный результат согласно потребностям пользователя.

Данная задача носит прикладной характер, так как при помощи матрицы инцидентности «термин-документ» становится возможным составлять инвертированные списки, которые являются основой полнотекстового информационного поиска. Таким образом, матрица инцидентности «термин-документ» является связующим звеном в процессе информационного поиска.

ЛЕКСИЧНО-СЕМАНТИЧНІ ОСОБЛИВОСТІ ТА ОСОБЛИВОСТІ ПЕРЕКЛАДУ ОКАЗІОНАЛІЗМІВ (НА МАТЕРІАЛІ РОМАНУ ДЖ. РОУЛІНГ "ГАРРІ ПОТТЕР І ОРДЕН ФЕНІКСУ")

Лебедєва Г.М.
Національний аерокосмічний університет ім. М.Є. Жуковського «ХАІ»
61070, м. Харків, вул. Чкалова, 17, тел.:(057) 3151131
e-mail:khai@khai.edu

Авторські новоутворення завжди були цікавим об’єктом досліджень, оскільки явище оказіональності є достатньо аномальним і суперечливим. З кожним роком лексичний рівень мови збагачується новими словами. Чи то новими технічними термінами, що пов’язані з новими винаходами, чи то дивними, незвичними словами та їх формами у художній літературі.

В нашій роботі ми розглядаємо оказіоналізм як об’єкт дослідження у лінгвістиці, як об’єкт змістового сприйняття і як об’єкт перекладу. Отже, для проведення цього дослідження ми використали 250 лексичних одиниць, вилучених з твору англійської письменниці Джоан Кетлінг Роулінг «Гарри Поттер и Орден Феникса» («Harry Potter and the Order of the Phoenix»). Семантика оказіоналізмів та способи їх перекладу стали предметом нашого дослідження.

Оказіоналізми (від лат. occasionalis – випадковий) − слова, що утворюються за наявними в мові моделями, але не використовуються в загальновживаному словнику. Оказіоналізми мають індивідуальний характер, вживаються тільки в умовах певного контексту, який дає змогу розкрити їхнє значення.

Повертаючись до предмету нашого дослідження, слід згадати головні способи перекладу оказіоналізмів. Отже, виділяють такі способи:

1) Транскрипція або транслітерація. При застосуванні цих способів перекладу відбувається акт запозичення звукової (транскрипція) або графічної (транслітерація) оболонки слова разом зі значенням із мови оригіналу до мови перекладу.

2) Калькування. Калькування як метод створення еквіваленту схожий на буквальний переклад. Перевагою цього методу є стислість і нескладність отриманого еквіваленту.

3) Описові еквіваленти. Цей метод полягає у переданні значення англійського слова за допомогою більш-менш розповсюдженого пояснення.

Практична мета нашої роботи полягала в аналізі лексичних одиниць за тематичним показником та за типом перекладу. Отже, за результатами семантичного аналізу було виділено такі групи: 1) Магічні тварини та рослини (Magical Creatures). Наприклад: *knarl – нарл, Umgubular Slashkilter – Чертохолопый Головосек* (12%); 2) Імена та прізвиська (Names and Nicknames). Наприклад: *Hedwig – Букля, Scabbers – Короста* (12%); 3) Заклинання (Spells) – найбільша група. Наприклад: *Expecto Patronum –Экспекто патронум*,

Invisibility Spell – чары невидимости (24%); 4) Приміщення та організації (Places and Organizations). Наприклад: *Azkaban – Азкабан, Gringotts – Гринготтс* (11%); 5) Зілля (Potions). Наприклад: *Invigoration Draught – Животворящий эликсир, Polyjuice Potion –Оборотное зелье* (5%); 6) Магічні речі (Magical things). Наприклад: *Howler – громовещатель, Invisibility Cloak – мантия-невидимка* (16%); 7) Їжа (Food). Наприклад: *Honeyduke – Сладкое королівство, Nosebleed Nougat – Кровопролитные конфеты* (6%); 8) Газеті та журнали (Newspapers and Magazines). Наприклад: *Sunday Prophet – Воскресный пророк, The Quibbler – Придира* (2%); 9) Гроші (Money). Наприклад: *Knuts– кнаты, Sickles– сикли* (1%); 10) Транспортні засоби (Transport). Наприклад: *Floo Network – Сеть летучего пороха, Hogwarts Express – Хогвартс-экспресс* (1%); 11) Події (Events). Наприклад: *Triwizard Tournament – Турнир трех волшебников, Yule Ball– Святочный бал* (1%); 12) Факультети Хогвартсу (Hogwarts Houses). Наприклад: *Hufflupeff – Пуффендуй, Slytherin – Слизерин* (2%); 13) Люди та їх сфера діяльності (People and Their Occupation). Наприклад: *Muggles – маглы, Occlumens – окклюменист* (5%); 14) Інше (Else). Наприклад: *Order of the Phoenix – Орден Феникса, Quidditch – квиддич*(2%).

Залежно від типу перекладу ми розподілили усі оказіоналізми між наступними групами: 1) Транскрипція. Наприклад: *Muggles – маглы, Occlumency – окклюменция* (5%); 2) Транслітерація. Наприклад: *Expelliarmus – Экспеллиармус, Finite – Финита* (15%); 3) Калькування – найбільша група. Наприклад: *Invisibility Cloak – мантия-невидимка, Mudblood – грязнокровка* (39%); 4)Описові еквіваленти. Наприклад: *Curse-Breaker – Ликвидатор заклятий, Dark Detector – Детектор Темных сил* (27%); 5) Змішаний тип. Наприклад: *Unspeakables – Невыразимцы, Veritaserum – сыворотка правды* (14%); 6) Дослівний переклад. Наприклад: *Death Eater – Пожиратель смерти, Devil's Snare – дьявольские силки.*

Проведене нами дослідження може слугувати підґрунтям для подальших досліджень у цьому напрямку та використовуватися на заняттях з практики перекладу як ілюстративний матеріал.

ІДЕНТИФІКАЦІЯ АНТРОПОНІМІВ У ПОВНОТЕКСТОВИХ ДОКУМЕНТАХ

Хайло А. М.
Національний технічний університет
"Харківський політехнічний інститут",
м. Харків, вул. Пушкінська, 79/2, тел. 707–63–60,
e-mail: *alina_khailo@ukr.net*

Автоматична обробка тексту на природній мові дозволяє полегшити пошук і вилучення інформації з метою подальшої аналітичної обробки. Здебільшого потрібен аналіз великих масивів коротких текстів з метою виділення значущої інформації. Автоматичне розпізнавання з подальшим виділенням в текстах власних назв, що позначають людей є особливою проблемою комп'ютерної обробки текстів на природній мові. Власні назви, на відміну від загальних назв, утворюють список, який постійно змінюється та доповнюється. Вирішення цього завдання пов'язане з проблемою ідентифікації власних назв та ланцюжків звичайних слів, які поводять себе в текстах як власні назви.

На сьогоднішній день не існує масштабних систем з автоматизованої обробки природньої мови, які здатні виокремлювати та маркувати антропоніми (за визначенням Виноградова В. С., антропонім – це власна назва (або набір назв), яка офіційно присвоєна окремій людині як її розпізнавальний знак) в текстах українською мовою. Крім цього, опису різноманітних систем ідентифікації антропонімів присвячена досить велика кількість зарубіжних публікацій. Їх автори пропонують різні методи розпізнавання та смислової інтерпретації, які можна умовно розділити на наступні групи:

• статистичний підхід (для створення статистичної моделі використовується корпус розмічених текстів, Sekine S., Eriguchi Y. [2000]);

• обчислювальні методи на основі навчальних моделей (наприклад, в рамках проекту CoNLL-2003);

• метод контекстного аналізу (спирається на правила ідентифікації антропонімів в тексті в залежності від лівого і правого контексту та списки слів відкритих класів, McDonald D. [1996], Kokkinakis D. [2004]);

• гібридний підхід (об'єднує статистичні методи і прийоми контекстного аналізу (Mikheev A. et al. [1998]).

У статті McDonald D. [1996] описується один з ключових компонентів системи розуміння природної мови Sparser – модуль PNF, який призначений не тільки для розпізнавання і класифікації власних назв, а й для виокремлення та інтерпретації антропонімічних груп.

Всього виділяються три етапи роботи, так чи інакше, пов'язаних з виділенням та обробкою антропонімів:

1) визначення меж послідовності слів, з яких утворюється антропонім;

2) віднесення отриманого елемента до тієї чи іншої семантичної категорії з одночасним дозволом неоднозначності;

3) збереження отриманого результату в моделі з метою його подальшого використання при роботі Sparser як з даними, так і з іншими текстами.

Показниками антропонімічних груп являються вже відомі системі власні назви, які зустрічаються в тексті, а також слова-класифікатори (наприклад, spokesman, company), які забезпечують можливість прогнозування: безпосередньо після таких лексичних одиниць велика ймовірність зустріти в тексті антропонім.

У статті Stevenson M., Gaizaukas R. [2000] обговорюється серія експериментів з системою, яка була побудована авторами на основі комплексу LASIE (різні його версії використовувалися в проектах MUC і HUB4). Метою цих експериментів була ідентифікація в текстах антропонімів на основі попередньо побудованих списків слів і словосполучень. Описувана авторами система представляє інтерес, оскільки вона є самонавчальною: користуючись досить простим набором фільтрів, програма поповнює раніше побудовані списки власних назв новими одиницями, в результаті чого вона стає набагато більш точною. У роботі викладаються способи побудови списків, їх поповнення, а також описуються фільтри і методи експериментальної роботи з ними.

Оскільки, на сьогоднішній день основна маса інформації зберігається і обробляється в електронному вигляді, практика показує, що більшість ділових пошукових завдань пов'язані з пошуком власних назв. Правильно виділяти і розпізнавати власні назви необхідно і при комп'ютерному аналізі текстів. До того ж, завдання вилучення антропонімів з тексту є критично важливою технологією для подальшого створення систем інформаційного пошуку і розуміння документів.

ТИПИ СЛОВНИКІВ: ТЕХНІЧНА ГАЛУЗЬ (СИСТЕМНИЙ ОГЛЯД)

Усов М.В.
Національний аерокосмічний університет ім. М.Є. Жуковського
«Харківський авіаційний інститут»
г. Харків, вул. Чкалова, 17, тел. 778-40-09,
e-mail: shtin_t@mail.ru

Словники – це невід'ємна частина процесу навчання, пізнавання світу, а також самоосвіти. *Об'єктом* нашого дослідження стали англомовні, україномовні та російськомовні словники технічної галузі: загального машинобудування та літакобудування. *Предмет дослідження* – принципи укладання словників та їх відповідність потребам зазначеної галузі, вивчення та укладання технічних словників та їх класифікація. *Метою* проведення дослідження є отримання знань про словники та їх повну класифікацію.

Лексикографія як наука є однією з наймолодших лінгвістичних дисциплін, адже свого наукового змісту набула лише у другій половині XX сторіччя. Довгий час цю науку вважали лише технікою укладання словників. Навколо велися дискусії щодо теоретичних питань та практики їх вирішення лексикографією.

Активного розвитку лексикографія спочатку набула в Англії. Перші словники починають видаватися вже на початку XVII сторіччя. Серед авторів слід відзначити вклад таких науковців як Роберт Кодрі, Н. Бейлі, Семюел Джонсон, Кенрік, Шерідан, Уонер. Американська лексикографія починає розвиватися дещо пізніше за англійську: Семюель Джонсон та Ной Вебстер. Словник останнього перевидавали близько 50 разів протягом багатьох років. Радянську лексикографію представила ціла низка видатних науковців: Б. А. Ларін, А. М. Бабкін, Г. В. Стєпанов, Ю. Н. Караулов, В. Д. Аракін, Л. В. Малаховський та ін.

Класифікувати словники вперше почали Л. Щерба, Г. Гак та Б. Городецький.

Сучасна класифікація словників стала розгалуженішою та повнішою, вона характеризує всі аспекти мовленнєвої діяльності людини. Якщо надати лише перелік існуючих на сьогодні різних видів словників (антропонімічні, діалектні (чи обласні), граматичні, словники сполучуваності, ідеографічні, семантичні й асоціативні словники, історичні, лінгвокраїнознавчі і культурологічні словники, морфемні і словотворчі, зворотні, орфографічні, орфоепічні, синонімічні, словники антонімів, лінгвістичних термінів, словники іноземних слів, неологізмів, омонімів, паронімів, скорочень, словники епітетів, порівнянь, метафор, словники соціальних і професійних діалектів, мови письменників і словники окремих творів, словники-довідники труднощів мови, термінологічні словники, тлумачні, топонімічні, етимологічні, фразеологічні та частотні словники), зможемо впевнитися, що словники відіграють велику роль у сучасній культурі: у них відбиваються знання, накопичені суспільством

протягом століть; вони служать цілям опису і нормалізації мови, сприяють підвищенню правильності і виразності мови його носіїв. Використовуючи різноманітні словники, людина може не тільки орієнтуватися в наукових та технічних спеціалізованих термінах, але за допомогою словників слів-синонімів збагачувати свою мову, робити її більш досконалою та цікавою для сприймання.

Укладання словника – доволі довгий процес, що потребує терпіння та послідовності у діях. Переклад термінів українською та російською мовами виявляється чи не найлегшою частиною проведення практичного дослідження для створення словника, адже зайняв найменше часу, а от пошук інформації у різних джерелах займає найбільшу кількість часу. Більшу частину інформації, потрібної для роботи, ми знайшли у мережі Інтернет на сайтах філологічного і навчального характеру.

Шкода, що в наш час, коли значення словників є настільки очевидним, так мало приділяється уваги навчанню школярів та студентів правильному користуванню різними видами словників. На наш погляд, це могло б слугувати росту їх мовної та пізнавальної культури.

Більшість словників (не тільки вузькоспеціалізованих для системних адміністраторів та перекладачів) переведені в електронні форми. Так, можна швидко знайти відомості про конкретного історичного персонажа, переглянути його фотографію чи навіть почути запис його голосу тощо. Мультимедійні словники, які сьогодні все більше і більше оточують сучасну людину, економлять час і місце. Сучасний комп’ютер тепер легко може вміщувати у собі не один десяток велетенських бібліотек, при чому пошук потрібної інформації настільки швидкий, що років 10-20 тому ці темпи видавалися б фантастичними.

Список літератури

1. *Баранов А. Н.* Введение в прикладную лингвистику. – М.: ЭдиториалУРСС, 2001. – 235с.

2. *Перебийніс В.І., Сорокін В.М.* Традиційна та комп'ютерна лексикографія / [навч. посібник]. – К., 2010. – 215с.

ОБЗОР МЕТОДОВ ФИЛЬТРАЦИИ СПАМА ЭЛЕКТРОННОЙ КОРРЕСПОНДЕНЦИИ

Варешнюк І. В.
Национальный технический университет "Харьковский политехнический институт",
г. Харьков, ул. Пушкинская, 79/2, тел. 707–63–60,
e-mail: *innominatus_tru@mail.ru*

Сейчас, в эпоху современных компьютерных технологий, спам является одной из основных проблем, с которой пользователям сети Интернет ежедневно сталкиваются, потому что они уже не в силах отказаться от возможностей, которые открыла для них всемирная паутина.

Под спамом в наиболее общем смысле понимают незапрашиваемую информацию, то есть, все, что Вы получаете, вне зависимости от Вашего желания.

Не зависимо от наличия коммерческих или рекламных целей спам всегда преследует одну цель – довести свою информацию до максимально возможного числа адресатов при минимальных издержках. Причем "авторов" не волнует состав аудитории, главное – количество [1].

Существует несколько основных видов спама, имеющих много общего и в то же время отличающихся друг от друга. Самым большим потоком спама является почтовый спам, который распространяется через электронную почту. В настоящее время доля вирусов и спама в общем трафике электронной почты составляет по разным оценкам от 70 до 95 процентов [2].

Поэтому в дальнейшем будем рассматривать именно этот вид спама [3].

Рассмотрим методы, используемые для фильтрации спама:

1. Статистические метода фильтрации спама. Программы автоматической фильтрации используют статистический анализ содержания письма для принятия решения, является ли оно спамом. На практике пользуются популярностью методы байесовской фильтрации спама или различные ее вариации.
2. Черные списки – списки IP-адресов компьютеров, о которых известно, что с них ведётся рассылка спама.
3. Авторизация почтовых серверов – различные способы для подтверждения того, что компьютер, отправляющий письмо, действительно имеет на это право (Sender ID, SPF, Caller ID, Yahoo DomainKeys, MessageLevel).
4. Серые списки. Метод серых списков основан на том, что «поведение» программного обеспечения, предназначенного для рассылки спама, отличается от поведения обычных почтовых серверов, а именно, спамерские программы не пытаются повторно отправить письмо при возникновении временной ошибки, как того требует протокол SMTP.
5. Проверка соблюдения требований протокола SMTP.
6. Общие ужесточения требований к письмам и отправителям, например –

отказ в приеме писем с неправильным обратным адресом (письма из несуществующих доменов), проверка доменного имени по IP-адресу компьютера, с которого идет письмо, и т. п.

7. Сортировка писем по содержанию полей заголовка письма даёт возможность избавиться от некоторого количества спама.

8. Системы типа «вызов-ответ» позволяют убедиться, что отправитель – человек, а не программа-робот. Использование этого метода требует от отправителя выполнения определённых дополнительных действий, часто это может быть нежелательно, так как многие реализации таких систем создают дополнительную нагрузку на почтовые системы.

9. Системы определения признаков массовости сообщения, такие как Razor и Distributed Checksum Clearinghouse.

10. Общее изменение идеологии работы электронной почты, при которой для принятия сервером получателя каждого сообщения система отправителя должна выполнить определенное «затратное» действие. Для обычных пользователей, отправляющих десятки писем, это не составит затруднения, тогда как затраты спамера умножаются на количество отправляемых им писем, обычно измеряемое миллионами [2].

В заключение, необходимо отметить, что каждый из вышеперечисленным методов, используемый по отдельности, является недостаточно эффективным.

Поэтому большинство коммерческих и свободно распространяемых спам-фильтров, используют одновременно несколько методов фильтрации, выстраивая их в цепочки, в которых команды SMTP и электронное письмо будет передаваться каждому фильтру по очереди. В случае отрицательного ответа хотя бы одного фильтра письмо будет отвергаться [4].

Список литературы

1. http://antispam.rin.ru/mytest2.htm
2. http://ru.wikipedia.org/wiki/Спам
3. http://www.securelist.com/ru/threats/spam?chapter=151
4. http://gate.udec.ntu-kpi.kiev.ua/~bat/exp/about-spam.html

ПІДБІР ТЕСТОВИХ ЗАВДАНЬ З ГРАМАТИКИ АНГЛІЙСЬКОЇ МОВИ ДЛЯ СТУДЕНТІВ ПЕРШОГО КУРСУ ФІЛОЛОГІЧНИХ ФАКУЛЬТЕТІВ (НА МАТЕРІАЛІ ГРАМАТИЧНИХ ТЕМ: THE PERFECT TENSES)

Охріменко Ю.М.
Національний аерокосмічний університет
ім. М.Є. Жуковського «ХАІ»
м. Харків, вул.Чкалова, 17, тел. 788-40-09,
e-mail: yuliya-okhrimenko@rambler.ru

Більшість дослідників, які працювали над вивченням питань тестування, прийшли до висновку, що тестовий контроль є ефективною формою контролю, яка відповідає цілям контролю, вимогам, що висуваються до нього, і забезпечує ефективну реалізацію всіх його функцій у процесі навчання іноземним мовам, саме це і визначає актуальність даної теми. Об'єктом дослідження нашого дослідження є тестовий контроль.

Предметом дослідження є тестовий контроль як одна з форм навчання іноземним мовам.

Метою нашого дослідження було дослідити та виділити особливості тестового контролю у навчанні іноземним мовам.

Під час дослідження нами були поставлені наступні завдання:

- ознайомитись з літературою з даної теми;
- виділити особливості тестування, як одного із засобів контролю вивчення іноземної мови;
- виділити особливості тестового контролю у процесі навчання іноземної мови;
- визначити та виділити проблеми тестового контролю;
- визначити та виділити методику складання тестів;
- розробити тестові завдання для тестового контролю студентів першого курсу філологічних факультетів (на матеріалі граматичних тем The Perfect Tenses);
- зробити відповідні висновки.

Навчання граматики англійської мови є невід'ємною частиною вивчення мови. Без начання граматики іноземної мови не може існувати нормативного говоріння та письма. Мета навчання граматичного матеріалу іноземної мови − це володіння граматичними навичками мовлення: репродуктивними, тобто граматичними навичками говоріння і письма (активним граматичним мінімумом) та рецептивними навичками, тобто граматичними навичками аудіювання і читання (активним і пасивним граматичним мінімумом) [1, с.14].

Застосування тестів під час проведення контролю засвоєного граматичного матеріалу доцільне тому, що вони задають напрям розумової діяльності тестованих, привчають їх варіювати процес переробки сприйманої інформації. Під тестом розуміються завдання, що мають специфічну організацію, яка дозволяє всім студентам чи учням працювати одночасно в однакових умовах і

записувати виконання символами [2, с. 12]. Основна відмінність тесту від традиційної контрольної роботи полягає у тому, що він завжди припускає вимірювання. Тому оцінка, що виставляється за підсумками тестування, відрізняється більшою об'єктивністю і незалежністю від можливого суб'єктивізму вчителя, ніж оцінка за виконання традиційної контрольної роботи, яка завжди суб'єктивна, оскільки заснована на враженні вчителя / викладача, не завжди вільного від його особистих симпатій або антипатій по відношенню до того або іншого учня / студента.

У даній роботі нами був створений банк тестових завдань, 320 прикладів, для перевірки граматичних навичок студентів першого курсу філологічних спеціальностей (на матеріалі граматичних тем The Perfect Tenses). Ці тести включають в себе перевірку різноманітних аспектів використання певного часу. До складу тестових завдань входять тести, мета яких − перевірити навички студентів в утворенні часів Present Perfect, Past Perfect, Future Perfect (по 50 тестових завдань на кожен час); тести, що направлені на закріплення навичок утворення питальних речень у часі Present Perfect (50 тестових завдань); більш ускладнені тестові завдання, в яких потрібно самостійно розкрити дужки та вибрати потрібний час Simple Past, Present Perfect (40 тестових завдань); тести на розкриття дужок та вибору необхідного часу серед Simple Past, Present Perfect, Past Perfect (40 тестових завдань) та тестові завдання, в яких потрібно правильно розкрити дужки, вибравши час Future Perfect або Future Simple (30 тестових завдань). Результати нашого дослідження можуть бути корисними в практичній діяльності студентів та викладачів іноземної мови.

Список літератури

1. *Афанасьева О. В.* Иностранные языки в школе / Ольга Васильевна Афанасьева // Методика преподавания английской грамматики − 1986. − №2. − С. 12−16.

2. *Фоломкина С.К.* Тестирование при изучении иностранного языка. / С.К. Фоломкина // Иностранные языки в школе − 1986. − №2. − С. 12−16.

МОДЕЛЬ ИНТЕЛЛЕКТУАЛЬНОЙ ОБРАБОТКИ ПАТЕНТНО-КОНЪЮНКТУРНОЙ ИНФОРМАЦИИ

Король О.И.
Национальный технический университет "ХПИ",
г. Харьков, ул. Фрунзе, 21, тел. 066 75 77 361,
e-mail: *korolka@bk.ru*

Современные справочно-информационные поисковые программы и патентные базы данных сохраняют детали того или иного объекта интеллектуальной собственности с довольно высокой точностью. Однако объемы данных растут, они легкодоступны в сети Интернет, но при этом не обладают точной и понятной структурой и не являются знанием в полном смысле слова. При обработке патентно-конъюнктурной информации (ПКИ) нужно получить закономерности, а не потоки и списки данных. Успех выполнения данных задач напрямую зависит от того, как быстро и качественно будет осуществляться процесс отбора информации из огромного массива не структурированных данных.

Предлагаем модель системы интеллектуальной обработки ПКИ, основанной на универсальном математическом аппарате – алгебры конечных предикатов (АКП) [1]. С помощью АКП могут быть описаны любые конечные отношения, что позволяет легко обнаружить и извлечь из входного корпуса термины и их свойства, отображать многоместные отношения, связывающее текстовую ПКИ.

Построенная модель разбивает каждый новый признак на не пересекаемые классы эквивалентности [2]. Она является полной, несократимой и не противоречивой. Ее часть приведена на рис. 1.

Прежде, чем представить системы предикатов ПКИ в виде понятном АКП, необходимо создать онтологии прикладной области, процесс построения которых состоит из обнаружения и извлечения терминов, реляционного анализа понятий и извлечения внешних отношений. В завершение, результаты выполнения данных двух этапов объединяются для получения более полной онтологии.

Путь к каждому предикату описывается формулами вида (1)-(2).

Объекты изобретения и полезной модели $X_{11}^{ИПМ}$:

$$\begin{cases} X_{11}^{Н} \vee X_{11}^{ПН} = 1, \\ X_{11}^{Н} \wedge \overline{X_{11}^{Н}} = 0, \\ X_{11}^{ПН} \wedge \overline{X_{11}^{ПН}} = 0. \end{cases} \qquad (1)$$

где $X_{11}{}^{Н}$ ($X_{11}^{\overline{Н}}$) – новое (не новое) изобретение,

$X_{11}{}^{ПН}$ ($X_{11}^{\overline{ПН}}$) – применение (не применение) по новому назначению.

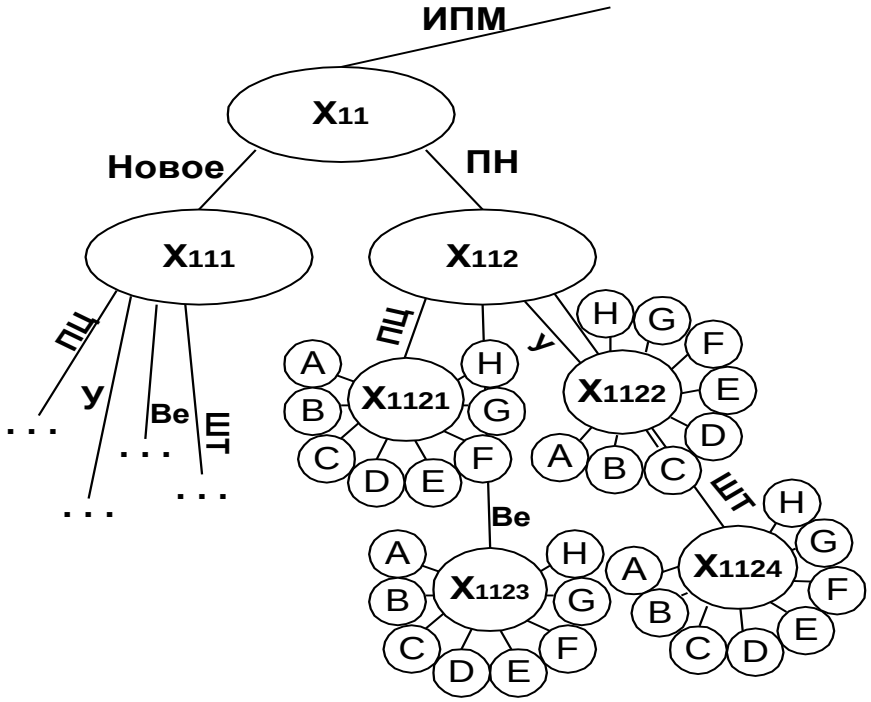


Рис. 1. Часть модели интеллектуальной обработки ПКИ.

ИПМ – изобретения и полезные модели, Новое – новое изобретение, ПН – применение ранее известного изобретения в новом качестве или новым способом, ПЦ – процесс, У – устройство, Ве – вещество, ШТ – штамм микроорганизма, A, B, C, D, E, F, G, H – классы изобретений по сфере применения.

Объекты нового изобретения $X_{111}^{ПН}$:

$$\begin{cases} X_{112}^{ПЦ} \vee X_{112}^{У} \vee X_{112}^{Ве} \vee X_{112}^{ШТ} = 1, \\ X_{112}^{ПЦ} \wedge X_{112}^{\overline{ПЦ}} = 0, \\ X_{112}^{У} \wedge X_{112}^{\overline{У}} = 0, \\ X_{112}^{Ве} \wedge X_{112}^{\overline{Ве}} = 0, \\ X_{112}^{ШТ} \wedge X_{112}^{\overline{ШТ}} = 0, \end{cases} \quad (2)$$

где $X_{112}^{ПЦ}$ - процесс, $X_{112}^{У}$ - устройство, $X_{112}^{Ве}$ - вещество, $X_{112}^{ШТ}$ - штамм микроорганизма, $X_{112}^{\overline{ПЦ}}$ - не процесс, $X_{112}^{\overline{У}}$ - не устройство, $X_{112}^{\overline{Ве}}$ - не вещество, $X_{112}^{\overline{ШТ}}$ - не штамм микроорганизма.

Список литературы

1. *Бондаренко, М.Ф.* Теория интеллекта [Текст] : учеб. / М.Ф. Бондаренко, Ю.П. Шабанов-Кушнаренко. – Харьков: Компания СМИТ, 2006. - 576 с.

2. *Король, О.І.* Технології побудови систем інтелектуальної обробки патентно-кон'юнктурної інформації [Текст] / О.І. Король // Вісник Херсонського нац. тех. ун-ту.– Херсон.– 2011.– №2 (41).– С. 163-165.

РОЗРОБКА ТЕСТОВИХ ЗАВДАНЬ З ПРЕДМЕТУ «ЛІНГВОКРАЇНОЗНАВСТВО» (ТЕМА: “HOLIDAYS, SYMBOLS AND TRADITIONS”)

Чалова С.Ю.
Національний аерокосмічний університет ім. М.Є. Жуковського "ХАІ"
м. Харків, вул.. Гв. Широнінців, 18/19, тел. 713-20-14
e-mail: lana_chalova@mail.ru, la_en@mail.ru

Знайомство з культурою країни, мова якої вивчається, є важливою частиною вивчення будь-якої іноземної мови, й такий предмет як лінгвокраїнознавство займається саме навчанням культурі країн через призму мови [1].

Критеріями відбору аутентичних матеріалів для навчання лінгвокраїнознавства є:

— культурна та країнознавча цінність;
— типовість;
— загальновідомість й орієнтація на сучасну дійсність;
— тематичність;
— функціональність [2; 3].

В Україні недостатньо аутентичних матеріалів та підручників з цього предмету. Також слід відзначити малу кількість якісних тестових завдань з лінгвокраїнознавства, а також їхню недостатню систематизованість. Це й визначає актуальність нашого дослідження.

Об’єктом дослідження слугували традиції та символи англомовних країн (зокрема, Великобританії та Сполучених Штатів Америки) [4].

Предметом є дослідження та відтворення тестових матеріалів з теми “Holidays, Symbols and Traditions”.

Метою дослідження є складення тестових завдань з теми “Holidays, Symbols and Traditions”.

Для досягнення поставленої мети були поставлені наступні завдання:

— ознайомитися з теорією складання тестових вправ;
— опрацювати матеріал з теми “Holidays, Symbols and Traditions in Great Britain and USA”;
— відібрати матеріали для тестових завдань;
— скласти тестові завдання, поділити їх на блоки за тематикою й групи за типом.

Матеріалом дослідження слугували підручники, конспекти, методичні посібники й довідкові матеріали з лінгвокраїнознавства, а також сайти Internet, які містили корисну інформацію на цю тему.

Тестовий контроль слугує для перевірки рівня знань, засвоєних студентами. Тестові завдання можуть використовуватися для домашнього завдання, для поточного та підсумкового контролю знань.

Наведені тестові завдання призначені для студентів спеціальності «Прикладна лінгвістика», що вивчають предмет «Лінгвокраїнознавство», а також усіх, хто вивчає цей предмет або цікавиться ним.

Через обмеження за кількістю виділеного для тестування часу, а також обмежень за часом, за який студенти можуть сконцентруватися на завданнях, ми включили до тесту два блоки по 33 завдання кожне.

Темою тесту є «Holidays, Symbols and Traditions». Ми розбили її на два блоки:

1. Блок «Свята, традиції й символи Великобританії»
2. Блок «Свята, традиції й символи ї Америки»

У кожному з блоків містилися завдання таких типів:

- завдання на встановлення відповідності;
- завдання множинного вибору;
- завдання вільного викладення;
- завдання альтернативних відповідей;
- завдання на заповнення пропусків (завдання-доповнення) ;
- кросворд.

Отримані тестові вправи можна використовувати на уроках, на екзаменах і на лінгвокраїнознавчих олімпіадах, для студентів та учнів старших класів середньої школи.

Список літератури

1. *Ахмаханова А. Е.* К вопросу об определении понятия «Лингвострановедение» [Електронний ресурс] // А. Е. Ахмаханова, З. Р. Тасанбаева – Академический Инновационный Университет, Шымкент. – Режим доступу до матеріалу : http://www.rusnauka.com/11_EISN_2010/Philologia/63965.doc.htm

2. Виды тестов [Електронний ресурс] // Softwerk. – Режим доступу до матеріалу : http://www.softwerk.ru/ttypes_r.htm

3. Виды тестирования [Електронний ресурс] // Режим доступу до матеріалу : http://luizaname.chat.ru/T/vidi.htm

4. *Гапонів А.Б.* Лінгвокраїнознавство. Англомовні країни. Підручник для студентів та викладачів вищих навчальних закладів. / А.Б.Гапонів, М.О. Возна – Вінниця: НОВА КНИГА, 2005. – 464 с.

ИСПОЛЬЗОВАНИЕ НАИВНОГО МЕТОДА БАЙЕСА ДЛЯ КЛАССИФИКАЦИИ КОЛЛЕКЦИИ ТЕКСТОВ

Лой А.А.
Национальный технический университет "Харьковский политехнический институт",
г. Харьков, ул. Пушкинская, 79/2, тел. 707–63–60,
e-mail: loy.alyna@gmail.com

Рост массивов полнотекстовых документов, публикуемых в интернете, требует новых средств организации доступа к информации. Одной из наиболее актуальных проблем управления знаниями, в особенности обеспечения быстрого информационного поиска в полнотекстовых базах знаний, является проблема автоматической классификации набора текстовых документов, которая представляет собой отдельный аспект задачи распознавания смысла текста.

Одним из эффективных алгоритмов классификации является так называемый «наивный» (упрощенный) алгоритм Байеса. Он основан на теореме, утверждающей, что если плотности распределения термов каждого из классов известны, то искомый алгоритм можно выписать в аналитическом виде. «Наивность» алгоритма заключается в предположении, что входные атрибуты условно (для каждого значения класса) независимы друг от друга.

Наивная байесовская модель является вероятностным методом обучения. Следуя предположению, что вероятности попадания термов в определенный класс независимы друг от друга, для получения вероятности в целом достаточно их перемножить. Вероятность того, что документ d попадёт в класс c, записывается как $P(c|d) \propto P(c)\prod_{1\le k\le n_d} P(t_k|c)$.

Здесь $P(t_k|c)$ – условная вероятность того, что терм t_k появится в документе из класса c (оценка вклада термина t_k в то, что документ принадлежит классу c), а $P(c)$ – априорная вероятность того, что документ принадлежит классу c. Последовательность $\langle t_1, t_2, \ldots, t_{n_d} \rangle$ состоит из значащих термов, а n_d – количество таких лексем в документе d. Поскольку цель классификации – найти самый подходящий класс для данного документа, то в наивной байесовской классификации задача состоит в нахождении наиболее вероятного класса c_m

В программной реализации обучающая коллекция представляет собой папку *Tutor*, содержащую набор подпапок, одноименных рассматриваемым классам. Динамически создаются таблицы, соответствующие классам и включающие два поля: индекс терма и его вес в рамках данного класса. Отдельно создаются таблицы *WORDS* (все термы обучающей коллекции) и *CLASSES* (информация о классах). Таблица *WORDS* содержит два поля: сам терм и его индекс; таблица *CLASSES* содержит три поля: название класса, его индекс и количество значащих термов в документах данного класса (без учета индивидуальности).

Чтение документов происходит посимвольно. Термом считается последовательность символов, ограниченных разделителями (все символы, кроме букв латинского алфавита, цифр и ‘&’). Каждый терм проходит морфологическую обработку (отсечение окончания, выявление неправильных глаголов, существительных множественного числа и т.п.). Числительные (последовательности символов, не содержащие букв) и отдельно стоящие буквы отсеиваются до этапа морфологической обработки и не считаются значимыми. Полученный терм сравнивается с таблицей стоп-слов. Стоп-словами считаются союзы, числительные прописью, местоимения, частицы и другие слова, не имеющие семантической значимости (междометия, вводные слова и т.п.). Если совпадений не найдено, терм считается значимым и информация о нем вносится в соответствующие таблицы.

Динамическое создание и заполнение таблиц позволяет реализовать простое редактирование обучающей выборки, а именно: добавление, удаление, объединение и переименование классов; добавление, удаление и редактирование документов, их перемещение в другие классы. Такая организация позволяет включить в интерфейс программы средства редактирования обучающей выборки пользователем.

После обработки документов обучающей коллекции, осуществляется классификация заданной пользователем коллекции. Для этого динамически создаются таблицы для каждого документа, состоящие из двух полей: индекса класса и вероятности $P(c|d)$.

Выделение термов происходит аналогично документам обучающей коллекции, вероятность $P(t_k|c)$ вычисляется для каждого значащего терма на основании информации таблицы *CLASSES* и его веса. К весу каждого терма прибавляется единица, чтобы избежать нулевой вероятности для не встретившихся в определенных классах термов. Термы, отсутствующие в таблице *WORDS* не учитываются. На основании созданной для документа таблицы определяется вероятность принадлежности данного документа к рассматриваемым классам.

Данный способ нашел широкое применение при фильтрации электронной почты, подбора контекстной рекламы, определении области поиска в поисковых системах, решении проблемы омонимии (полисемантичности) слов, что особенно важно для решения задачи автоматизированного перевода.

LINGUISTIC AND CULTUROLOGICAL ASPECTS OF ENGLISH-LANGUAGE COMMUNICATION IN THE FIELD OF EDUCATION

Shevchenko Anna
M. Ye. Zhukovsky National Aerospace University
Kharkiv Aviation Institute
Chkalova str., 17, Kharkiv,
Email: annschewtschenko@gmail.com

The intercultural communication in the field of education is assuming special importance in view of Ukraine's integration into the Bologna process. Consequently, the issue of advanced study of educational systems of the countries we contact with and permanently growing practical needs in intercultural communication required investigation of interrelations between the language and the culture.

The terminology in the field of education practically has not been investigated in the national science of terminology in the contrastive aspect. One of the prospective tasks of the national lexicography is to compile electronic English-Ukrainian and Ukrainian-English dictionaries of terminology in the field of education that would contain explanations of realia new to our language and culture. This is precisely why the works exploring the academic communication retain high priority.

There are different methods of creation of the English terms in the field of education. They are frequently created according to the metaphorization model. The vocabulary of the field of education can also be created through affixation (**pretest**). The process of abbreviation is intensive in the modern English. The following abbreviations can be distinguished in the English pedagogical term-system: syllabic abbreviation (**bach**<bachelor), initial abbreviations **(CBE –** Computer-based education), acronyms (**MACOS** – Man: A Course of Study), telescopic words (**Oxbridge** < Oxford + Cambridge).

However, by no means all English terms in the field of education widely used in the English scientific texts and academic communication are presented in English-Ukrainian and English-Russian dictionaries. This results from differences in language realia in cultures of the countries.

A language reflects the culture serviced by it in its vocabulary. The meaning of a word can be different depending on the culture it is servicing. In linguistics and psycholinguistics, the "lacunae", as is customary, are understood as basic elements of the national specificity of a linguocultural community that impair understanding of some text fragments by recipients from other cultures.

Though the term "lacuna" is used only in the cases referring to the absence of some concept in the target language, which cannot be explained, all lacunae should not be treated as non-equivalent lexis. Realia can be recognized as the closest to lacunae, since they denote a notion absent in the target language. Realia as objects of the material and spiritual culture reflect the way of life and thinking of a particular society and have no analogues in the other culture, the language of which also has no lexical units denoting unknown cultural concepts.

The educational activity is just the type of activity for which very illustrative are diversities in connotations and realia of different countries. That is why we can observe here a great number of lacunae and non-equivalents words, not fixed in Ukrainian dictionaries.

In our work, we have analysed 260 lexical units in several aspects. For more convenient analysis, the units were divided by the following topical groups: equipment and documents, scores and evaluation, courses, programs, classes, personalities; public organizations, entrance and registration, academic degrees and types of diplomas, an educational institution's structure and constituent parts, educational institutions, education financing, tasks and methods of study, process of study.

Thus, the following diagram reflects the situation concerning lacunar and non-equivalent lexis: lacunae — 28%, non-equivalent lexis — 25%, common language realia — 57%.

Regarding such vocabulary availability in Ukrainian dictionaries, one can say that quite a great number of terms are not fixed in modern translation dictionaries. All lexical units can be divided into units with translation available in English-Ukrainian and English-Russian dictionaries (68%), and those with no lexicographically fixed translation (32%).

The simplest ways to translate non-equivalent lexis are transliteration and transcription. However, traditionally such method is not productive in translation of the vocabulary of the field of education. The method of loan translation is widely used in translation of educational terminology. An English word can also be replaced with an already existing (synonymic) equivalent, which is at this stage used less frequently than a "trendy" English word.

Thus, the following results show the methods the most frequently applied for translation of the lexis under consideration: loan translation — 16%, descriptive translation — 22%, synonymic equivalent — 62%.

The results of the conducted analysis of the terminology of the field of education revealed a tendency towards development of a relevant term-system in the modern Ukrainian language. However, unlike others, the educational terminology is not adequately investigated by Ukrainian scientists. From the point of view of lexicography, translation of many frequently used English terms is missing in Ukrainian translation dictionaries. Lacuna and non-equivalent lexis demands particularly thorough examination and adequate reproduction in lexicographic sources, including electronic dictionaries.

ВЫЯВЛЕНИЕ СЕМАНТИЧЕСКИХ ЭКВИВАЛЕНТОВ ПРИ АВТОМАТИЧЕСКОЙ ОБРАБОТКЕ ТЕКСТОВ

Петрасова С.В.
Национальный технический университет "Харьковский политехнический институт",
г. Харьков, ул. Пушкинская, 79/2, тел. 707–63–60,
e-mail: svetapetrasova@gmail.com

Целью работы является анализ существующих методов выявления семантических эквивалентов и выбор метода для решения задачи их автоматического определения в заданном текстовом ресурсе экономической направленности.

В настоящее время существует множество методик информационного поиска. Все они могут быть поделены на три большие группы: статистические методы поиска, методы поиска по семантическим сетям и комбинированные методы поиска.

Семантический поиск это метод поиска, в котором релевантность документа запросу определяется с использованием семантических, а не статистических методов, как происходит в подходах информационного поиска по ключевым словам.

Семантические эквиваленты – текстовые выражения, сопоставленные одному и тому же понятию. Семантическими эквивалентами являются синонимы и семантические близкие слова. Под семантически близкими словами подразумеваются слова с близким значением, встречающиеся в одном контексте.

В качестве базовых знаний для автоматической семантической обработки, обычно используются онтологии, которые представляют собой формальные описание терминов предметной области и отношений между ними, и тезаурус, как особая разновидность словарей общей или специальной лексики, в которых указаны семантические отношения (синонимы, антонимы, паронимы, гипонимы, гиперонимы и т.п.) между лексическими единицами. Таким образом, тезаурусы, особенно в электронном формате, являются одним из действенных инструментов для описания отдельных предметных областей. Благодаря использованию онтологий и тезаурусов удается строить образ достаточно релевантный запрашиваемому. Этот образ может использоваться для формирования более эффективных запросов для поисковой системы.

Из статистических моделей поиска остановимся на следующих. В "моделях векторных пространств" формируются векторные представления слов и других компонент текстов путем автоматического извлечения статистики их совместной встречаемости из больших массивов текстовой информации.

В основе метода латентно-семантического анализа лежит гипотеза о том, что между отдельными словами и обобщенным контекстом (предложениями, абзацами и целыми текстами), в которых они встречаются, существуют

неявные (латентные) взаимосвязи, обуславливающие совокупность взаимных ограничений.

Метод Клейнберга (HITS алгоритм) использует понятия авторитетного и хаб-документа и заключается в анализе ссылок, что позволяет ранжировать веб-страницы. Авторитетный документ – это документ, соответствующий запросу пользователя, имеющий больший удельный вес среди документов данной тематики. Хаб-документ – это документ, содержащий ссылки на авторитетные документы. Метод основан на вычислении собственного вектора матрицы, описывающей структуру ссылок в вебе.

Наиболее перспективной является группа методов, которые объединяет качественную статистическую модель поиска и учет семантических конструкций. К этой группе можно отнести метод расстояний, который положен в основу исследования. Он использует в качестве лингвистического ресурса толковый словарь, который позволяет дать количественную оценку семантической близости между терминами словаря. Суть метода расстояний заключается в том, что два слова считаются синонимами, если имеют общие слова в своих определениях (понятиях).

Для реализации метода осуществляется следующий алгоритм: выполняется предварительная лингвистическая обработка; создается лингвистическая база данных, состоящая из терминов и их определений; после введения пользователем запроса проводится попарное сравнение термина запроса с каждым термином словаря; степень семантической близости определяется расстоянием между двумя терминами, которое представлено в виде суммы количества компонент в определении первого термина, отсутствующих в определении второго, и количества компонент в определении второго термина, отсутствующих в определении первого. Величину расстояния нормализуем путем отношения полученной суммы необщих компонентов к сумме всех компонент первого и второго термина. В результате данного сравнения пользователь получает набор терминов, имеющих общие компоненты с термином запроса.

Предлагаемое данным методом решение задачи автоматического выявления семантических эквивалентов может использоваться в поисковых системах для расширения запроса, для автоматизированного построения онтологии по тексту, для расширения существующих и создания новых тезаурусов.

СЕМАНТИЧЕСКИЙ АНАЛИЗ КОНЦЕПТОВ «ЛЮБОВЬ» В РУССКОМ И «LOVE» В АНГЛИЙСКОМ ЯЗЫКАХ

Коняева К. Г.
*Национальный технический университет
"Харьковский политехнический институт",
г. Харьков, ул. Пушкинская, 79/2, тел. 707–63–60,
e-mail: par4ik@mail.ru*

Важнейшее достижение современной лингвистики заключается в том, что язык уже не рассматривается отдельно от других наук, он представляется в новой парадигме с позиции его участия в познавательной деятельности человека. Знания об окружающем мире, которые человек получает, структурируются в языке в процессе коммуникации. Главные единицы языковой картины мира человека называются концептами.

Концепт — это содержательная сторона словесного знака, за которой стоит понятие, относящееся к умственной, духовной или материальной сфере существования человека, закрепленное в общественном опыте народа, имеющее в его жизни исторические корни, социально и субъективно осмысляемое и, через ступень такого осмысления, соотносимое с другими понятиями, с ним связанными [1].

Проанализировав понятия «любовь» в русском и «love» в английском языке, можно сделать вывод о частичном различии этих понятий даже на уровне их определений в толковых словарях.

Из семантических значений ключевой лексемы можно выделить следующие концептуальные признаки исследуемого концепта: любовь может быть самоотверженной (любовь к родине); любовь может быть инстинктивной (материнская любовь); любовь может быть основана на половом влечении (чувственная любовь); любовь может проявляться в форме склонности, влечения к чему-либо (любовь к искусству) [2]. Часть значений, приведенных английским толковым словарем (любовь как половое влечение, любовь как склонность к чему-либо) совпадает с русским вариантом, однако английский словарь не выделяет признак самоотверженности любви [3]. Кроме того, семантика лексемы love не отражает признак материнской любви, как одной из ее форм, хотя в английских дефинициях достаточно обобщенно упоминается любовь к членам семьи и друзьям.

Концепт имеет сложную структуру, но в общем виде ее можно представить в виде поля, в центре которого лежит основное понятие – это ядро концепта, а на периферии находится то, что привнесено в это понятие культурой, традициями, народным и личным опытом [1].

И в русском, и в английском языковом сознании присутствуют и являются ядерными такие концептуальные признаки как истинность и бескорыстие любви, крайне обостренное чувство любви, что сближает ее с пристрастием. Однако интерес представляет тот факт, что концептуальный признак

способности любви причинять страдания является периферийным в английском языковом сознании, его рейтинг составляет всего 2%, и ядерным в русском языковом сознании, где рейтинг выше в 3 раза - 6% (муки любви, несчастная любовь, от любви страдать, умирать, изнемогать).

Концепт можно охарактеризовать его лексико-семантическим полем, которое представляет собой иерархическую организацию слов, объединенных одним родовым значением и представляющих в языке определенную семантическую сферу, которая покрывает некоторую область действительности [4].

В результате сопоставления лексико-семантических полей, в составе которых находилась базовая лексема «любовь» и «love», был сделан вывод о том, что множество микрополей совпадают. Например, микрополе «чувство по отношению к противоположному полю», «самоотверженная привязанность к кому-либо или чему-либо» и т. д. При этом смысловая нагрузка перечисленных полей в различных лингвокультурах может различаться. Ярким примером служит микрополе «пристрастие к чему-либо». В английском языке положительным пристрастием считается пристрастие к машинам, литературе, музыке, драгоценностям, качественной одежде. В русском языке положительным является пристрастие к комфорту. К числу отрицательных в английском языке относится пристрастие к оружию, а в русском – к спиртному. Следовательно, состав положительных и отрицательных пристрастий в данных языках не имеет сходства, поскольку его специфика отражает исторически сложившиеся вкусы носителей сравниваемых языков.

Проведенное исследование дает основания делать некоторые объективные выводы о менталитете, приоритетах носителей языка по национальным признакам, что в очередной раз доказывает значимость исследований в данной области.

Список литературы

1. *Лингвистический энциклопедический словарь*. Гл. ред. В. Н. Ярцева. М.: Научное изд-во «Большая Российская Энциклопедия», 2002.

2. *Ожегов С. И., Шведова Н. Ю.* Толковый словарь русского языка: 80 000 слов и фразеологических выражений. — 4-е изд., М., 1997. — 944 с.

3. *Cambridge International Dictionary of English*: — Москва, Cambridge University Press, 2001 г.

4. *Маслова, В.А.* Когнитивная лингвистика: Учебное пособие / В.А.Маслова.– Мн.: ТетраСистем, 2004.

ИСПОЛЬЗОВАНИЕ ОНТОЛОГИЙ ДЛЯ ПРЕДСТАВЛЕНИЯ ПРЕДМЕТНОЙ ОБЛАСТИ

Иванющенко В.С.
Национальный технический университет "Харьковский политехнический институт",
г. Харьков, ул. Пушкинская, 79/2, тел. 707–63–60,
e-mail: *toshe4ka1104@yandex.ru*

Онтологии являются новыми интеллектуальными средствами для поиска ресурсов в сети Интернет, новыми методами представления и обработки знаний и запросов. Они способны точно и эффективно описывать семантику данных для некоторой предметной области и решать проблему несовместимости и противоречивости понятий.

Онтология – формальная спецификация разделяемой концептуализации, которая имеет место в некотором контексте предметной области. При этом под концептуализацией будем иметь в виду, кроме сбора понятий, также всю информацию, касающуюся понятий – свойства, отношения, ограничения, аксиомы и утверждения о понятиях, необходимые для описания и решения задач в избранной предметной области.

По степени зависимости от конкретной задачи или предметной области различают:

- онтологии верхнего уровня;
- онтологии, ориентированные на предметную область;
- онтологии, ориентированные на конкретную задачу;
- прикладные онтологии.

Все модели онтологий, в той или иной степени, содержат концепты (понятия, классы, сущности, категории), свойства концептов (слоты, атрибуты, роли), отношения между концептами (связи, зависимости, функции) и дополнительные ограничения (определяются аксиомами, в некоторых парадигмах фасетами)

Semantic web - это направление развития web-технологии, целью которого является представление информации в виде, пригодном для машинной обработки.

Методологическая модель RDF - важная компонента Semantic web, назначение которой состоит в описании отношений между сетевыми ресурсами и информацией. RDF представляет собой технологию для выражения смысла терминов и понятий в виде, доступном для обработки программами. Эта технология предназначена для стандартизации определений и использования метаданных, описывающих Web_ресурсы, а также для представления самих данных, содержащихся в этих ресурсах.

RDF использует базовую модель данных «объект — атрибут — значение» и способен сыграть роль универсального языка описания семантики ресурсов и взаимосвязей между ними. Ресурсы описываются в виде ориентированного

размеченного графа — каждый ресурс может иметь свойства, которые в свою очередь также могут быть ресурсами или их коллекциями. Все словари RDF используют базовую структуру, описывающую классы ресурсов и типы связей между ними

Для разработки онтологии была выбрана область IT-технологий.

Информационные технологии (от англ. information technology, IT) — широкий класс дисциплин и областей деятельности, относящихся к технологиям создания, управления и обработки данных, в том числе с применением вычислительной техники.

Основные черты современных ИТ:

- компьютерная обработка информации по заданным алгоритмам;
- хранение больших объёмов информации на машинных носителях;
- передача информации на значительные расстояния в ограниченное время.

В работе был проанализирован способ построения онтологий предметной области, рассмотрен язык написания онтологии и выбрана предметная область для её разработки.

Список литературы

1. *Гаврилова Т. А., Хорошевский В. Ф.* Базые знаний интеллектуальных систем. – С.–Пб.: Питер, 2001

2. *Клещев А. С., Артемьева И. Л.* Математические модели онтологий предметных областей. Часть 1. Существующие подходы к определению понятия «онтология». // Научно-техническая информация, серия 2 «Информационные процессы и системы», 2001, № 2, с. 20–27.

3. *Клещев А. С., Артемьева И. Л.* Математические модели онтологии предметной области. Часть 2. Компоненты модели. // Научно-техническая информация, серия 2 «Информационные процессы и системы», 2001, № 3, с. 19–28.

МЕТОДЫ ОБРАБОТКИ СТРУКТУРИРОВАННЫХ ДОКУМЕНТОВ

Ляхвацкая О.Н.
Национальный технический университет
"Харьковский политехнический институт",
г. Харьков, ул. Пушкинская, 79/2, тел. 707–63–60,
e-mail: *belsquirrel213@mail.ru*

Большинство исследователей под обработкой естественно-языкового текста традиционно понимают обработку текста, которая представляет собой набор предложений без выраженной структуры. В настоящее время становится актуальной задача обработки документов, обладающих высокой степенью формальности и, как следствие, внутренней иерархической структурой [1].

Сложность задачи анализа иерархически структурированных текстов обусловлена следующими их свойствами:

1) как правило, разметка заголовков и маркеров (с помощью стилей, тегов и т.д.) в документе присутствует лишь частично или отсутствует;

2) заголовки с различных уровней иерархии могут не отличаться по виду;

3) название и ссылка в тексте могут иметь одинаковый вид;

4) большое количество конфигураций непрерывных текстовых фрагментов: предложение может состоять из нескольких таких фрагментов, один фрагмент может включать несколько предложений, группа предложений может быть вложена в предложение в виде комментариев [1].

Документы можно различать по стилю форматирования:

1. Структурированные документы, в которых расположение и размеры полей фиксированы.

2. Частично структурированные документы, для которых известен перечень реквизитов, но не регламентировано их расположение и количество.

3. Гибко-структурированные документы или «гибкие документы» – документы, в которых состав и порядок следования их частей по горизонтали и вертикали одинаков, но части могут отличаться по размерам или масштабу.

4. Свободно структурированные документы, у них нет обязательных реквизитов, и форматирование не ограничено [2].

Исходя из видов структурированных документов, рассматриваются такие методы обработки (Data Mining):

1. Алгебраические методы. Исходные данные в них представляются в виде алгебраических структур.

2. Статистические методы. Они используют аппарат теории вероятностей и математической статистики.

3. Методы мягких вычислений. В них используются нечеткое представление данных (нейросети, генетический алгоритм и т.д.) [3].

В данной работе подробно рассматриваются такие виды структурированных документов, как патенты. Патенты – это документ, который свидетельствует о праве изобретателя на его изобретение.

Преимущества системы патентования:

- поощряет изобретателей изобретать;
- обнародование помогает другим исследователям;
- после окончания срока - вседозволенности.

Недостатки:

- опасность монополии.

В работе проанализированы поля данных патентов, рассмотрено из чего состоит патент. Проведено сравнительный анализ патентов Национального Украинского Патентного Бюро с Американским Патентным Бюро и найдено ряд отличий. Для применения и создания программы решено использовать методы Data Mining.

Применение методов DM оправдано при наличии достаточно большого количества данных, которые имеются в структурированных документах. А так же методы DM позволяют доступно и понятно провести парсирование документов, в данном случае патентов.

Список литературы

1. *Лахути Д.Г.* Автоматический анализ естественно-языковых текстов. // М.: ВИНИТИ, 2003.

2. *Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И.* Методы и модели анализа данных: OLAP и Data Mining.

3. *Шереметьева С.О.* Теоретические и методологические проблемы инженерной лингвистики. // М.: ВИНИТИ, 1998.

ЛЕКСИКОГРАФИЧЕСКИЙ АСПЕКТ ОПИСАНИЯ ЛЕКСИЧЕСКИХ ПАРАЛЛЕЛЕЙ

Цыбизова Ю. С.
Национальный технический университет «Харьковский политехнический институт», г. Харьков, ул. Фрунзе 21, тел. (057) 700-15-64, e-mail: omsroot@kpi.kharkov.ua

Выступление посвящено краткому анализу словарей и теоретических положений ученых, которые исследовали тему внешне сходных слов в различных языках, которые могут полностью/частично/неполностью совпадать по значению.

1. Первым словарем, который я проанализировала, был «Англо-русский и русско–английский словарь “ложных друзей переводчика”» В.В.Акуленко (М., 1969)[1]. **Межъязыковые синонимы -** это слова двух языков, полностью или частично совпадающие по значению и употреблению (и, соответственно, являющиеся эквивалентами при переводе). **Межъязыковыми омонимами** можно назвать слова обоих языков, сходные до степени отождествления по звуковой (или графической) форме, но имеющие разные типы значения. Наконец, к **межъязыковым паронимам** следует отнести слова сопоставляемых языков, не вполне сходные по форме, но могущие вызвать у большего или меньшего числа лиц ложные ассоциации и отождествляться друг с другом, несмотря на фактическое расхождение их значений.

2. Второй словарь, описывающий данную лексику, - «Немецко-русский и русско-немецкий словарь “ложных друзей переводчика”» К.Г.М.Готлиба (М., 1972) [2]. Рассматриваются **междуязычные аналогизмы** с точки зрения их внешней структуры, он определяет степень их фоно-морфологической, графической и семантической близости. устанавливает два типа:

3. Эту же проблему затронул проф. М.П.Кочерган, создавая «Словарь русско-украинских межъязыковых омонимов» (К., 1997) [3]. **Межъязыковые омонимы** — это полностью или в основной части совпадающие по форме, но различающиеся содержанием слова двух контактирующих языков. Например: рус. *неделя* «семь дней, седмица», укр. *неділя* «воскресенье».

4. Проанализировав все эти словари, я решила, что лучше всего опереться на термин “лексические параллели”, предложенный в 1993 г. проф. В.В.Дубичинским [4]. Лексемы, совпадающие в плане выражения и сходные/несходные в плане содержания называются этим обобщающим термином “**лексические параллели**”. Если внешне сходные лексемы сравниваемых языков семантически полностью совпадают, то такие лексические параллели называются **полными.** Например, укр. *архітектура* и нем. *Architectur*. В случае совпадения одних и несовпадения других значений семантических структур внешне сходных лексем речь идет о **неполных лексических параллелях**. Например укр. *диктат* и нем. *Diktat*. В данной классификации совпадающие значения принято называть **интерсемемами**, а

несовпадающие, отражающие национально-культурное своеобразие лексической единицы - **идиосememами.** Понятия интерсемем и идиосемем дает возможность провести сопоставительный (переводческий) анализ на уровне отдельных значений лексем и в определенных случаях иногда учесть даже семантические и стилистические нюансы на уровне более мелких компонентов значений – сем. В случае же несовпадения всех значений внешне похожих лексических единиц двух или более синхронически сравниваемых языков речь идет о **ложных лексических параллелях.** Например, укр. *актор* и нем. *Akteur*.

Данные теоретические положения будут продемонстрированы в докладе на примерах словарных статей из **Украинско-немецкого словаря лексических параллелей**, который создается сейчас в творческом сотрудничестве лингвистов Харьковского лексикографического общества при НТУ «ХПИ» (Украина) и Института славистики Клагенфуртского университета (Австрия).

Список литературы

1. *Акуленко В.В.* Англо-русский и русско-английский словарь “ложных друзей переводчика” - М., 1969.

2. *Готлиб К.Г.М.* Немецко-русский и русско-немецкий словарь “ложных друзей переводчика” - М., 1972.

3. *Кочерган М.П.* Словник російсько- українських міжмовних омонімів - К., 1997.

4. *Дубичинский В.В.* Лексические параллели – Харьков, 1993; *Дубичинский В.В., Ройтер Т.* Русско-немецкий словарь лексических параллелей – М., 2011.

ПРОЕКТУВАННЯ МУЛЬТІМЕДІЙНОЇ НАВЧАЛЬНОЇ СИСТЕМИ ДЛЯ ВИВЧЕННЯ АНГЛІЙСЬКОЇ МОВИ

Прогляда Я. В.
Национальный технический университет "Харьковский политехнический институт",
г. Харьков, ул. Пушкинская, 79/2, тел. 707–63–60,
e-mail: *Yana-Proglyada@yandex.ru*

Нинішній етап розвитку суспільства характеризується швидкими темпами розповсюдження інформації. Інформаційні технології використовують в усіх галузях людського життя. Необхідною умовою застосування інформаційних технологій є реформування системи освіти, розробка нових дидактичних і методичних концептуальних засад освіти.

У зв'язку з розвитком процесу інформатизації і освіти змінюється об'єм і зміст учбового матеріалу, відбувається реорганізація програм учбових предметів (курсів), інтеграція самих учбових предметів, що призводить до зміни структури і змісту учбових предметів і, отже, структури і змісту освіти. Навчання з використанням інформаційних технологій якісно перевищує класичну освіту, інтегруючи процеси, які не можна об’єднувати в межах класичної освіти.

В даний час вже створено безліч засобів навчання за допомогою комп'ютера. Їх можна кваліфікувати наступним чином: комп'ютерні підручники, предметно-орієнтовані середовища (мікросвіти, моделюючі програми, навчальні пакети), лабораторні практикуми тренажери, контролюючі програми. У наш час існує небагато мультимедійних комплексів, які присвячені вивченню саме граматики англійської мови.

Дана робота направлена на вдосконалення розробки мультимедійного комплексу для інтерактивного вивчення граматики англійської мови.

З точки зору принципів сприйняття інформації при спілкуванні за допомогою комп'ютера виділяють, як правило, два теоретичних підходи: біхевіористичний і когнітивно-інтелектуальний.

Біхевіористичний підхід, пов'язаний з постулатом «чим частіше вжито слово, тим краще воно запам'ятовується», в останні роки удосконалився використанням цілого ряду прийомів: дедуктивного контролю відповідей, створенням універсальних банків даних, побудовою довідкової інформації у вигляді гіпертекстів та інше. Суть методу інформування полягає в тому, що в пам'ять комп'ютера розміщуються деякі довідково-інформаційні дані (граматичний довідник, орфографічний словник, двомовний словник і т.п.), користувач може використовувати їх при підготовці до занять чи у процесі занять. Проте бихевіористичний підхід не може подолати механістичність навчання і відсутність розвитку когнітивних здібностей учнів. При когнітивно-інтелектуальному підході у користувача активізуються пізнавальні функції. Для успішної реалізації такого підходу в пам'яті комп'ютера створюється

універсальне навчальне середовище, що включає в себе різні граматичні довідники, словники, системи перевірки орфографії, інші допоміжні матеріали.

При розробці навчальної системи «Easy English» було вибрано когнітивно-інтелектуальний підхід. Було сформульовано загальні принципи побудови та використання комп'ютерних навчальних систем. Проаналізовано існуючі системи інтерактивного вивчення іноземної мови, визначено їх переваги та недоліки. Розглянуто існуючи можливості проектування змісту курсу навчальної системи та його складових. Проведено підготовку навчального матеріалу. Розроблено начальні сценарії такі як: утворення ступенів порівнянь прикметників і утворення множини іменників англійської мови. Створено програмну оболонку для майбутньої мультимедійної програми.

В якості програмної оболонки було використано систему управління змістом Joomla, , написана на мовах PHP і JavaScript, яка використовує як сховища бази даних MySQL. Джумла має наступні переваги: управління веб - посиланнями; управління даними в різних медіа форматах (.jpg, .gif, .bmp, .png і flash); управління отриманням новин з різних спеціалізованих сайтів; управління користувачами сайту з встановленням різних видів доступу; доступ до статистики відвідувань.

Розробка мультимедійного комплексу для інтерактивного вивчення граматики англійської мови є дуже важливою і корисною роботою, та допоможе тим, хто вивчає іноземну мову опанувати всі аспекти вивчення мови. Її використання призведе до швидкого вивчення нової лексики та граматики. Після кожної пройденої теми система оцінює рівень засвоєння нового матеріалу користувачем та вказує на його помилки, що є дуже важливим при вивченні іноземної мови.

Список літератури

1. *Карлащук В.И.* Обучающие программы. – М., 2001. – 330 с.
2. *Носенко Э. Л.* Применение ИТ в образовании. – М., 1990. – 256 с.
3. *Башмаков А.И., Башмаков И.А.* Разработка компьютерных учебников и обучающих систем – М.: Информационно-издательский дом «Филинъ», 2003. – 616 с.

ПРОБЛЕМЫ И ОСОБЕННОСТИ ПОСТРОЕНИЯ ИНВЕРТИРОВАННОГО ИНДЕКСА КОЛЛЕКЦИИ ДОКУМЕНТОВ

Луда С. Э.
Национальный технический университет
«Харьковский политехнический институт»
г. Харьков, ул. Херсонская, д.12, кв.1
+380937276595, e-mail: *simplement.sy@gmail.com*

Информационный поиск может осуществляться по разным алгоритмам, но именно алгоритм инвертированных индексов сегодня используется всеми крупными поисковыми системами в мире. В чем же заключается процесс индексации коллекции документов?

При использовании алгоритма обратных индексов, поисковые системы преобразовывают документы в инвертированные текстовые файлы. Каждый такой файл представляет собой индексную структуру, состоящую из двух частей:

1 – словарь, содержащий термины, дополнительные структуры, обеспечивающие быстрый поиск по термину;

2 – пост-листы. Каждый пост-лист представляет собой массив адресов вхождений слова - идентификатор документа, или идентификатор документа и позиций слова в документе, или дополнительные флаги и форматирования слова и т.п. Каждый список словопозиций упорядочен по идентификаторам документа.

В ходе анализа был разработан прототип системы для индексации коллекции документов на языке C++ в среде «Borland C++ Builder». Он производит индексацию трех текстов: «Document Indexing Tutorial.txt», «Introduction to IR.txt», «IR Wikipedia.txt». Рассмотрим алгоритм построения инвертированного индекса этой коллекции.

На первом этапе программа размечает текст, с помощью функции strtok() разбивая его на лексемы - последовательности символов, объединенные в семантическую единицу для обработки.

На следующем этапе осуществляется нормализация лексем – это процесс приведения лексем к канонической форме, осуществляемый для устранения несущественных различий между последовательностями символов. В результате мы получаем список терминов. Термин – это нормализированная лексема, включенная в словарь системы информационного поиска.

В данной программе все лексемы приведены к нормализованному виду терминов вручную, так как в нашем примере достаточно разделить текст по пробелам и отбросить знаки пунктуации. Однако при обработке большой коллекции документов возникает необходимость решения более сложных задач, таких как установление функций дефиса или апострофа, а также пробела, т.к. некоторые слова, разделенные пробелом, семантически обозначают одну лексему (New York, Los Angeles). Некоторые системы игнорируют т.н. стоп-

слова (stop-words) - распространенные слова, не представляющие ценности для удовлетворения информационных потребностей пользователей, например, a, and, be, for, has, he, is, it, to, и др. Однако при поиске фразы ее смысл может быть утерян, если не индексировать эти слова. Так, например, некоторые фразы целиком состоят из стоп-слов («To be or not to be», «Let it be»).

Часто в поисковых системах игнорируется регистр букв (case-folding). Но это может привести к непреднамеренному расширению запроса, т.к. многие имена собственные отличаются от имен нарицательных лишь прописной буквой. Примером являются General Motors, The Fed(Федеральная резервная система), фамилии Bush, Black.

Следующая проблема состоит в том, что по грамматическим причинам в документе встречаются разные формы одних и тех же слов. Существует два способа решения – это стемминг и лемматизация. *Стемминг* – это процесс, в ходе которого от слов отбрасываются окончания с расчетом на то, что в большинстве случаев это себя оправдает. *Лемматизация* – это точный процесс с использованием лексикона и морфологического анализа слов, в результате которого удаляются только флективные окончания и возвращается основная форма слова, называемая *леммой*. Например, лексема «saw» в ходе стемминга может превратиться в букву «s», а лемматизация вернет либо слово «see», если эта лексема является глаголом, либо «saw», если она – имя существительное. Несмотря на то, что для некоторых процессов лемматизация может оказаться очень полезной, для остальных запросов она существенно снижает производительность. Стемминг же повышает полноту, но снижает точность поиска.

На последнем этапе программа индексирует документы, в которых встречаются термины, создавая инвертированный индекс, состоящий из словаря и словопозиций. В результате мы получаем таблицу, в которой все термины, содержащиеся в коллекции, расположены в алфавитном порядке. Напротив каждого термина указана частота, с которой он встречается в каждом из документов коллекции отдельно и его общая частота во всей коллекции.

Структура обратных индексов подобна глоссарию книги, в котором указано, где найти документ. Сама идея инвертированного индекса практически не имеет конкурентов, поскольку является наиболее эффективной для текстового поиска по произвольному запросу.

СПЕЦИАЛИЗИРОВАННЫЙ КУРС КАК СРЕДСТВО ФОРМИРОВАНИЯ КОМПЕТЕНТНОСТИ

Поморцева Е. Е.
Харьковский национальный экономический университет
Харьков, пр. Ленина, 9-а, e-mail: elena_pomor@rambler.ru

Рассмотрены вопросы, связанные с восполнением пробелов в знаниях по информационным технологиям у студентов первого курса. Использование для этих целей курсов «Университетское образование» и «Выравнивающий курс по информатике» позволит решить данную проблему в течение первого учебного семестра.

В мировом сообществе бурно развиваются процессы информатизации и компьютеризации всех сфер деятельности человека. От уровня информационно-технологического развития и его темпов зависит состояние экономики и качество жизни людей. Информатизация образования предусматривает изменение содержания, методов, организационных форм и технологий обучения, оснащение учебных заведений компьютерной техникой, пересмотр учебно-методического обеспечения образовательных программ, повышение квалификации преподавателей, административных и инженерно-технических кадров. Компьютеры, информационные технологии не только пронизывают все технические дисциплины, в том числе и точные науки. Они меняют и сами эти дисциплины, и методику их преподавания.

Для того чтобы студент смог успешно осваивать материал необходимо вводить различные специализированные курсы. Необходимым условием формирования информационной культуры студентов, является преемственность содержания школьного и вузовского образования, в частности, информатики. Для поддержания этой преемственности на первом курсе всех специальностей были введены две новые дисциплины – «Университетское образование» и «Выравнивающий курс по информатике». Именно в такой последовательности. Это связано, прежде всего с тем, что в ходе преподавания дисциплины «Университетское образование» студентам должны быть изложены основные правила и способы работы с университетским сайтом дистанционного обучения.

Использование студентами материалов в электронном виде по дисциплинам предполагает, во-первых, наличие у студентов компьютера, подсоединенного к Интернету, и во-вторых, знание ими компьютерных технологий обработки и обмена данными. Как правило, эти первоначальные знания студенты получают только при изучении на первом курсе дисциплины «Информатика». Поэтому возникает потребность упреждающего обучения студентов базовым знаниям по информатике, чтобы они с первых дней учебы могли использовать сайт дистанционного обучения университета. Именно в этих целях и была введена дисциплина «Университетское образование». Кроме того, буквально на первых занятиях дисциплины «Университетское образование» было проведено входное тестирование по школьной информатике

Рис. 1 – Распределение оценок при проведении входного тестирования.

студентов первого курса. По результатам этого тестирования формируются группы на «Выравнивающий курс по информатике». Результаты показали крайне низкие остаточные знания. Это вело бы, с одной стороны, к невозможности воспользоваться теми материалами, которые выложены на сайте дистанционного обучения университета, а с другой стороны, к большому проценту отсева студентов по результатам первой экзаменационной сессии.

Понять, как составить программу выравнивающего курса помогают, с одной стороны, тесты, которые прошли студенты, с другой стороны, контрольная работа, которую пишут все студенты группы на самом первом занятии данной дисциплины. Результаты обработки тестов, которые проходят все без исключения студенты показывают, что знания по информатике оставляют желать лучшего (рис.1). Целью курса «Выравнивающий курс по информатике» является ликвидация того пробела в знаниях, который необходим студенту для дальнейшей успешной учебы. Ведь понятно, что в настоящее время, особенно учитывая специфику экономического вуза, слабая компьютерная подготовка будет причиной проблем при освоении большинства предметов на старших курсах и специальных дисциплин, при написании курсовых и дипломных работ. Особенностью организации учебного процесса по данному курсу является то, что по нему отсутствуют лекции, а предусмотрены только лабораторные занятия. Именно то, что по дисциплине «Выравнивающий курс по информатике» предусмотрены только лабораторные работы [1] является одним из достоинств и позволяет в течение одного семестра восполнить пробелы в знаниях. При выполнении лабораторных работ с помощью преподавателя студент может овладеть теми компетенциями, которые необходимы для получения умений и навыков использования компьютерных технологий в предметной области будущего специалиста.

Для того чтобы убедиться в пользе проведения курса «Выравнивающий курс по информатике», в конце семестра было проведено тестирование студентов с использованием тех же тестов, что и в самом начале семестра. Результаты второго тестирования приведены на рис. 2.

Рис. 2 – Распределение
оценок после предложенной формы организации занятий.

Таким образом, необходимо чтобы «Выравнивающий курс по информатике» оставался полноценной отдельной дисциплиной со своими целями и задачами. Только тогда этот курс пойдет на пользу, как студенту, так и вузу в целом. Доля студентов, отчисленных по неуспеваемости после первой сессии значительно снизится и студент, имевший по тем или иным причинам пробелы в знаниях, не только восполнит их, но и будет чувствовать себя наравне со всеми остальными студентами. То есть проблем, связанных с приобретением компетенций, как по дисциплинам компьютерного цикла, так и по дисциплинам на старших курсах у него не будет.

Список литературы

1. *Выготский Л.С.* Собр. соч.: в 6 т. Т. 3. М.: Педагогика, 1983. – 674 с.

ОСОБЛИВОСТІ ФУНКЦІОНУВАННЯ НАЙЧАСТОТНІШОЇ ЛЕКСИКИ В АНГЛОМОВНОМУ АКАДЕМІЧНОМУ СПІЛКУВАННІ

Скапа Л.В.
Національний аерокосмічний університет ім. М.Є. Жуковського
Харківський авіаційний інститут
Харків, вул. Чкалова, 17
Email: kustistaya@mail.ru

Дуже актуальним в наш час є питання міжнародного співробітництва. Це й не дивно, адже Україна має висококваліфікованих фахівців, в чиїх дослідженнях та розробках зацікавлені іноземні колеги. На жаль, небагато науковців які займаються дослідженнями у галузі техніки, володіють англійською мовою на рівні, що дозволяє їм вільно спілкуватися з іноземними колегами. Та що необхідно знати українському науковцю для того, щоб адекватно викладати свої науково-теоретичні міркування та результати практичних досліджень? На це питання ми намагалися дати відповідь у ході нашого дослідження.

Науковий дискурс – багатожанрове функціональне утворення. Слід зазначити, що, яким би не був той чи інший жанр наукового дискурсу, з якою метою він би не з'являвся, учасники наукового дискурсу вирішують вужчі, конкретніші завдання, а саме: 1) отримання нової наукової інформації (обмін думками, листування науковців); 2) знаходження вірного рішення (творче обговорення, мозковий штурм); 3) відстоювання своєї або критику чужої думки (наукова полеміка); 4) фіксацію проміжних або остаточних результатів своєї дослідницької діяльності (стаття, монографія).

Для того, щоб не почуватися незручно під час міжкультурної комунікації, необхідним є знання моделей спілкування, культурних стереотипів, ціннісних орієнтирів, образів і символів культури. Треба знати, як носій мови сприймає власну культуру, розуміти норми вербальної та невербальної поведінки, уміти моделювати свою поведінку з урахуванням цих особливостей та норм під час контактів з носіями мови; володіти національно-специфічними формами спілкування, мовними та поведінковими кліше.

Лексика є дуже нестабільною стороною будь-якої мови. Сьогодні всі застосовують той чи інший термін, а завтра він уже вийшов з ужитку. Задля того, щоб уникнути використання рідковживаних мовленнєвих одиниць, нами було створено частотний словник усіх відібраних нами слів за матеріалами, які знаходяться на сайті корпусу сучасної американської англійської мови (CORPUS OF CONTEMPORARY AMERICAN ENGLISH). На сьогодні даний корпус містить 410 мільйонів слів та великий масив текстів різних жанрів, він існує у вільному доступі, однак потребує реєстрації. За допомогою інструментів цього корпусу можна з'ясувати, скільки разів зустрічалося необхідне слово в академічних виданнях в період з 1990 по 2011 роки. Наприклад, такі іменники, як *research, data, approach, science, evidence, addition, issue, project, report,*

review, є найуживанішими в науковому стилі, а українському науковцю, який хоче добре володіти англійською мовою, треба починати саме з вивчення найуживанішої лексики. Аналіз, проведений за допомогою корпусу, надає достовірну інформацію про комбінаторні характеристики (сполучуваність) досліджуваних одиниць. Так, наприклад, іменник *evidence* у значенні „доказ" в академічному стилі вживають набагато частіше, ніж *proof*. Ця інформація є корисною як для науковців, які користуються англійською мовою, так і для перекладачів.

Взагалі тема функціонування англійської лексики в науковому співробітництві є дуже актуальною та все ще потребує детальнішого вивчення.

ЗАСТОСУВАННЯ МУЛЬТИМЕДІЙНОГО ОБЛАДНАННЯ НА ЗАНЯТТЯХ З ГРАМАТИКИ АНГЛІЙСЬКОЇ МОВИ

Кулєшова Т.І.
Національний аерокосмічний університет ім.М.Є. Жуковського
«Харківський авіаційний інститут»
м. Харків, вул. Чкалова, 17, тел. 788-40-09
e-mail: ella1301@gmail.com

Останнім часом все частіше постає питання про використання мультимедійних технологій у процесі навчання. Спеціально розроблені мультимедійні засоби дозволяють зробити процес навчання більш доступним, зрозумілиим; такий спосіб пізнання нової інформації заохочує учнів/студентів до вивчення нового матеріалу та закріплення вивченого. Зважаючи на це, наше дослідження є *актуальним* – граматика будь-якої мови (в нашому випадку – англійської) із застосуванням мультимедійних технологій на заняттях надає матеріалові новизни та здатна зацікавить молодь у вивченні. Отже модернізація методів навчання, без сумніву, приносе свої позитивні результати.

Докладне вивчення статтей, наукових праць, підручників, журналів та спеціальних посібників, присвячених дослідженню, створенню та використанню інформаційних технологій та засобів мультимедіа на заняттях з вивчення граматики англійської мови (а саме – розробка навчальних фільмів, презентацій, створення тестів тощо), надало нам можливість розробити серію допоміжних матеріалів з граматики англійської мови для студентів першого курсу ВНЗ філологічних спеціальностей, які застосовуються на заняттях з використанням новітніх засобів мультимедіа – електронних дошок, графічних та звукових засобів впливу, компьютерних програм з вивчення англійської мови тощо.

Зазначимо, що під терміном “*multimedia*” розуміють інтерактивні системи, що забезпечують обробку рухомих і нерухомих відеозображень, анімованої графіки, високоякісного звука та мовлення [1, с.25].

Мультимедійні технології мають безперечні переваги над іншими навчальними технологіями. Це – по-перше, активізація освітнього процесу за рахунок посилення наочності. А оскільки інформація є дуже складною субстанцією, яку людина сприймає згідно своїм здібностям та використовує у різних сферах свого життя, мультимедійні технології активізують інтерактивну взаємодію, що дозволяє управляти пред’явленням інформації, надають можливість поєднання логічного та образного способів засвоєння інформації, а також вивчення результатів, можливість змінювати швидкість подання інформації та кількість повторювань, що повинні задовольняти індивідуальні академічні потреби [2, с.40–41].

Оптимальним засобом мультимедія можна назвати створення мультимедійної Power Point презентації. Застосування комп'ютерних презентацій на заняттях дозволяє ввести новий лексичний, граматичний або

країнознавчий матеріал у найбільш захоплюючій формі, оскільки реалізується принцип наочності, що сприяє міцному засвоєнню інформації. Самостійна творча робота студентів зі створення комп'ютерних презентацій як найкраще розширює запас активної лексики та влаштовує перевірку знань граматики під час підбору та створення тестових завдань.

На сучасному етапі навчання студентів надзвичайно поширеним методом підсумкової або проміжної перевірки є тестування. Залежно від можливостей викладача, студентам пропонується пройти тестування на роздрукованих бланках, на персональних комп'ютерах в аудиторії, локальній мережі або у мережі Інтернет. Під час роботи на комп'ютері можлива обробка матеріалу за допомогою спеціальних програм для складання тестів (Test Designer).

Відповідно до *об'єкту* та *основних цілей* дослідження нами було розроблено чотири презентації та три відеозаписи для використання на заняттях з граматики англійської мови для студентів першого курсу на основі графічного матеріалу. Для їх створення ми підібрали ілюстрації для більш наглядного зображення прикладів. Ми використали приклади речень (і надали їм анімаційний графічний супровід), що ясніше відображають певний граматичний матеріал, необхідний для вивчення, як, наприклад, *«Let's dance»* та *«Don't let him speak»* за темою «Наказовий спосіб дієслова в англійській мові». Ці презентації підвищують якість навчального процесу й ефективність засвоювання матеріалу та надалі можуть бути використані викладачем на заняттях для вивчення зазначеного граматичного матеріалу.

Окрім створення елементарних презентацій для подання матеріалу, сучасний розвиток комп'ютерної техніки надає можливість використовувати легкі навчальні фільми, створені на основі презентації.

Освоївши просту програму Windows Movie Maker, кожен бажаючий може створити короткий навчальний фільм, куди він зможе включити не тільки музику або рухомі зображення, а також свої власні коментарі, прочитані у голос.

Список літератури

1. *Лактіонов О. Б.* Мультимедія – новій напрям комп'ютеризації освіти/ О.Б. Лактіонов // Рідна школа, 1993. – №3. – С. 25.

2. *Ушакова С.В.* Компьютер на уроках английского языка/С.В. Ушакова//ИЯШ. [Педагогические науки], 1997. – №5. – с. 40–41.

КОМПЬЮТЕРНАЯ ИГРА «ВЕЛИКИЕ ЛЕКСИКОГРАФЫ»

Данилевич С. Б.
Харьковский гуманитарный університет
«Народная украинская академия»,
г. Харьков, ул. Лермонтовская 27, тел. 714-20-07,
e-mail: profkom@nua.kharkov.ua

Использование в учебном процессе интеллектуальных игр – это один из путей повышения эффективности познавательного процесса. Так, известный педагог Сухомлинский В. А. писал: «Игра – это искра, зажигающая огонек пытливости и любознательности» [1].

В тезисах описывается методика и практические рекомендаций по созданию деловых игровых программ средствами MS Excel с применением VBA. За основу взята деловая (некомпьютерная) игра, описанная в [2].

Целью игры является усиление мотивации к изучению учебной дисциплины Лексикография.

Во время игры обучаемому на экране компьютера представлены цитаты выдающихся лексикографов. Необходимо определить их авторов. Цитаты взяты из книги [3].

Чтобы выбрать автора предложенной цитаты нужно щелкнуть по надписи с фамилией автора левой кнопкой мыши или перетащить ее мышкой при нажатой левой кнопке. В случае успешного выбора появляется надпись с поздравлением и новая цитата, выбираемая программой случайным образом. В противном случае из списка авторов исчезает неудачно выбранная фамилия и фиксируется номер попытки. Так как игра имеет целью ознакомить обучаемого с мыслями, афоризмами, взглядами выдающихся людей, то предусмотрена подсказка. При небольшой модификации игровая программа может быть преобразована в тестирующую.

Список цитат и авторов помещен на листе Excel (см. таблицу 1).

Таблица 1.

11	**Л.Н. Толстой**	**Слово есть поступок.**	**Время проходит, но сказанное слово остается.**	**Все мысли, которые имеют огромные последствия, всегда просты.**

В дальнейшем тексты можно изменить. В редакторе VBA можно спроектировать диалоговое окно (рис. 1).

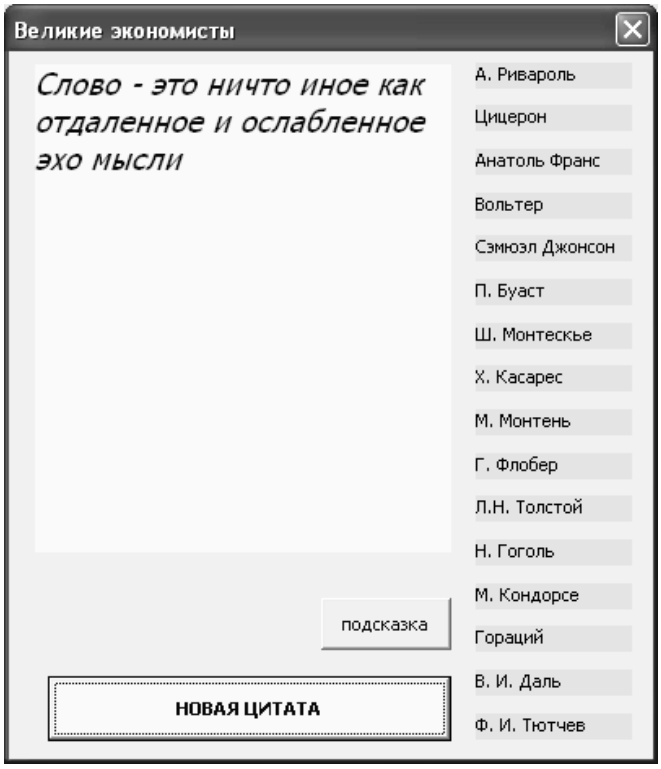


Рис. 1.

Цитата выбирается случайным образом. Выбор автора цитаты осуществляется щелчком мыши.

Список литературы

1. *Сухомлинский В. А.* Сердце отдаю детям. – Киев: Радянська школа, 1969. – 381с.

2. *Корнейчук Б.В.* Микроэкономика. Деловые игры. – СПб.: Питер, 2003. – 157 с.

3. *Дубичинский В.В.* Теоретическая и практическая лексикография. Wiener Slawistischer Almanach, sonderband 45, Wien-Charkov, 1998. – 160 с.

ОСНОВНІ ПРОБЛЕМИ ВИКОРИСТАННЯ МАШИННОГО ПЕРЕКЛАДУ ДЛЯ ПЕРЕКЛАДУ ХУДОЖНІХ ТВОРІВ

Періжняк М.М.

Національний технічний університет
«Харківський політехнічний інститут»
м. Харків, вул. Командарма Корка,буд.38, кв.31
Тел: 067-7252145. E-mail: k4spian@gmail.com

На сьогоднішній день машинний переклад – це цілий спектр технологій, що швидко розвиваються і мають широке поле застосування. Проте, існують окремі види перекладацької діяльності, використання машинного перекладу у яких не є ефективним або доцільним. Саме таким видом є переклад художніх творів.

Переклад художніх творів вимагає від перекладача не тільки знання мови оригіналу і навичок перекладу, а також глибокого рівня фонових знань, вміння працювати зі стилем і абстрактного мислення. На сучасному етапі розвитку науки і техніки, всі ці вимоги важко перенести на комп'ютерну програму або веб-застосунок, що виконує функції системи машинного перекладу. В результаті, машинний переклад у цій сфері не застосовується загалом або застосовується тільки як допоміжний засіб, при цьому основні функції виконує людина-перекладач. Проте, із збільшенням об'ємів інформації і вдосконаленням технологій її машинної обробки, ростуть запити щодо швидкості усіх видів перекладу, в тому числі і перекладу художніх творів. В цьому контексті, робота над створенням системи машинного перекладу, що могла б ефективно виконувати такі функції, є актуальним завданням прикладної лінгвістики.

Перш ніж досліджувати перспективи створення такої системи, необхідно окреслити основні проблеми або труднощі, що не дозволяють сучасним системам машинного перекладу ефективно працювати із художніми творами. А саме:

1. Нездатність систем машинного перекладу ефективно перекладати художні засоби, такі як метафора, гра слів, іронія та інших. Особливо, якщо ці засоби є авторськими.
2. Труднощі перекладу фразеологізмів та крилатих виразів. Особливо тих, що не є широко вживаними.
3. Проблема передачі авторського стилю у перекладі художніх творів.
4. Необхідність зробити текст перекладу милозвучним і зрозумілим для читача як принципова вимога до перекладу художніх творів.
5. Важливість збереження художньої цінності твору.

Кожна з цих проблем заглиблюється коренями у більш загальну проблему – проблему суб'єктивності художнього мовлення. Остаточне вирішення цих проблем є завданням майбутнього, проте рух у цьому напрямку має починатися вже сьогодні.

ИСПОЛЬЗОВАНИЕ ОНТОЛОГИЙ ДЛЯ РЕШЕНИЯ ЗАДАЧ OPINION MINING

Терещенко В.И.
Национальный технический университет
«Харьковский политехнический институт»,
г. Харьков, ул. Фрунзе 21,
тел. (057) 700-15-64,
e-mail: omsroot@kpi.kharkov.ua

Сегодня все большую популярность в развитии семантических сетей и автоматическом анализе текстовой информации приобретает такое направление как Opinion Mining. Не смотря на то что этим вопросом занимается очень большое количество лингвистов и программистов, задача оценки мнения текста остается до конца нерешенной. В связи с тем, что каждый день в интернете накапливается все больше информации в текстовом виде относительно того или иного явления или продукта, появилась потребность в средствах быстрого анализа информации и получения основных данных, представляющих наиболее сжатую и в то же время, информативную выжимку из текста. Например при покупке телефона, поиск информации в интернете об определенной модели и существующих о ней мнений может занять от нескольких часов до нескольких дней. Однако даже с такими затратами времени и сил, пользователь не может быть до конца уверенным в своем решении поскольку полученная информация является не полной и ограничивается несколькими десятками сайтов или форумов. Методы и технологии Opinion Mining являются более эффективными, поскольку позволяют проанализировать большие объемы информации из множества различных источников (от нескольких сотен до нескольких тысяч) за короткий срок времени [1].

Для релевантного определения мнения текста, технологии Opinion Mining наряду с задачей определения мнения текстов, должны включать в себя задачу обработки семантических эквивалентов. Поиск семантических эквивалентов неразрывно связан с задачами Opinion Mining, поскольку он позволяет находить в тексте слова с эквивалентным значением [2]. Хотелось бы отметить, что частично эти задачи решаются с помощью онтологий и тезаурусов [3].

Предметом данного исследования являются онтологии и тезаурусы с помощью которых в дальнейшем стало бы возможным приблизиться к решению проблем Opinion mining.

В ходе исследования были проанализированы такие online-библиотеки онтологий как “DAML ontology library”, “Protégé ontologies library”,“Code4Lib library ontology”, содержащие большое количество онтологий разных направлений. После проведенного анализа, для дальнейшего исследования были выбраны следующие онтологии: WordNet, Ruthes, Gemet, SentiWordNet.

В рамках исследования проведен сравнительный анализ этих онтологий, связей используемых в них а также проанализирована возможность их

использования для решения задачи Opinion Mining и поиска семантических эквивалентов. В результате исследования получены следующие выводы:

— в онтологии SentiWordNet, наилучшим образом определены связи используемые для решения задачи Opinion Mining;

— в онтологии WordNet определены связи для поиска семантических эквивалентов.

Однако эти онтологии не решают поставленные задачи полностью. SentiWordNet как и WordNet можно использовать только для слов и некоторых словосочетаний, однако остаются почти бесполезными для анализа текстовой информации, как то новостные рассылки или другие её источники.

В святи с этим, был сделан вывод о том что для решения задач Opinion Mining и поиска семантических эквивалентов наиболее действенным будет разработка собственного тезауруса или онтологии на основе существующих и с использованием основных принципов и методов их построения.

Список литературы

1. *Большакова Е.И., Баева Н.В., Васильева Н.Э.* Структурирование и извлечение знаний, представленных в научных текстах / Девятая Национальная конференция по искусственному интеллекту с международным участием КИИ-2004. Труды конференции в 3-х томах. Т. 2. - Москва: Физматлит, 2004. – 480-488 с.

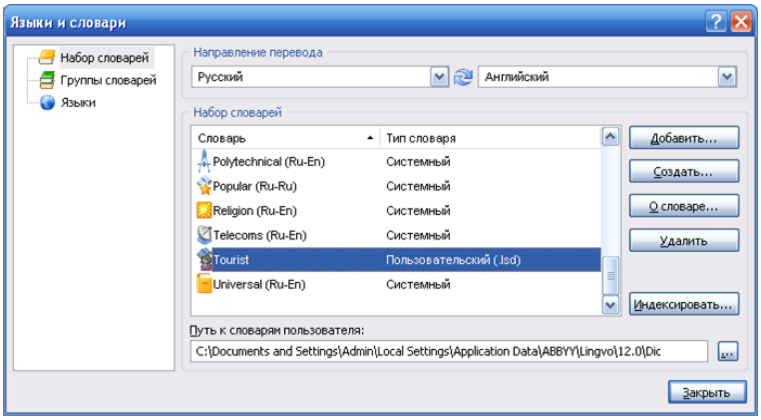
2. *Пэнг Б., Ли Л.* Извлечение мнений и смысловой анализ. Т.2. / Б.Пэнг, Л. Ли. - Foundations and Trends in Information Retrieval., 2008.-135 с.

3. *Соловьев В.Д., Добров Б.В., Иванов В.В., Лукашевич Н.В.* Онтологии и тезаурусы: Учебное пособие / В.Д. Соловъев, Б.В. Добров, В.В. Иванов, Н.В. Лукашевич. – Казань: Москва, 2006. – 157 с.

РАЗРАБОТКА ЭЛЕКТРОННОГО ПЕРЕВОДЧИКА НА ПЛАТФОРМЕ МАШИННОГО ПЕРЕВОДА APERTIUM

Васильева Ю.В.
Харьковский гуманитарный университет
«Народная украинская академия»,
г. Харьков, ул. Лермонтовская, 27, тел. 704-10-37,
e-mail: yuljav90@mail.ru

В условиях современной информационной среды все популярней становится участие пользователей в работе над различными интернет-проектами. Этот процесс получил свое развитие благодаря переходу на новую концепцию Сети, получившую название Web 2.0. Суть этой концепции состоит в возможности и эффективности работы над проектами путём совместной деятельности пользователей. Ярким примером таких проектов являются системы перевода и системы словарей, которые открыты для разработки и пополнения.



Разработанный нами в предыдущие годы русско-английский электронный словарь «Tourist», интегрированный в оболочку программы Lingvo в ХГУ «НУА», может быть с успехом применен в словарных системах, доступных широкому кругу пользователей, с целью использования для перевода как отдельных слов и фраз, так и для применения в системе машинного перевода. Для дальнейшего использования разработанного нами словаря мы выбрали платформу машинного перевода Apertium (http://www.apertium.org), в которой есть возможность размещения словарей в процессе создания системы машинного перевода для различных языковых пар. На данной платформе создано уже около 40 автоматических переводчиков и ещё многие находятся в процессе создания. Команда Apertium проявляет большой интерес к региональным языкам и активно поддерживает работы по созданию новых систем автоматического перевода, для которых разрабатывается лингвистическая база из словарей и правил в чётко указанных форматах XML. С этой целью используется простой XML-стандарт на основе форматов для кодирования необходимых языковых данных [1].

Система Apertium представляет собой платформу машинного перевода поверхностно-трансферного типа. Данная платформа имеет дело со словарями и правилами поверхностного трансфера. Правила представляют собой операции с группами лексических единиц, входящих в три словаря: два морфологических

словаря для языков рассматриваемой пары, содержащих информацию о словоизменении (склонении или спряжении) на каждом из языков (apertium-rus-en.rus.dix и apertium-rus-en.en.dix), и двуязычный словарь, который содержит переводные соответствия слов и символов двух языков (apertium-rus-en.rus - en.dix) [2].

```
<section id="main" type="standard">
<e><p><l>автобус<s n="n"/></l><r>bus<s n="n"/></r></p></e>
<e><p><l>троллейбус<s n="n"/></l><r>trolleybus<s n="n"/></r></p></e>
<e><p><l>стол<s n="n"/></l><r>table<s n="n"/></r></p></e>
<e><p><l>шкаф<s n="n"/></l><r>cupboard<s n="n"/></r></p></e>
<e><p><l>аромат<s n="n"/></l><r>aroma<s n="n"/></r></p></e>
<e><p><l>телефон<s n="n"/></l><r>phone<s n="n"/></r></p></e>
<e><p><l>балкон<s n="n"/></l><r>balcony<s n="n"/></r></p></e>
```

Кроме того, составляющими языковую пару являются также два файла с правилами трансфера. Правила трансфера русского языка на английский язык описывают, каким изменениям подвергнутся предложения русского языка при переводе на английский язык (apertium-rus-en.rus -en.t1x).Правила трансфера английского языка на русский язык описывают преобразования, которые должны быть осуществлены при переводе с английского языка на русский язык (apertium-rus-en.en- rus.t1x) [2]. Следует отметить, что часто такие языковые ресурсы, как корпуса, словари, грамматики, морфологические анализаторы, списки лемм, находятся в свободном доступе либо с лицензией на возможность свободного использования, и могут быть использованы повторно, что значительно сокращает время разработки нового переводчика.

Таким образом, учитывая имеющийся опыт работы в данном направлении, можно утверждать, что идея создания русско-английского переводчика на базе платформы машинного перевода Apertium является достаточно актуальной и перспективной сферой деятельности в области развития современных языковых компьютерных технологий.

Список литературы

1. Apertium [Электронный ресурс]: – Режим доступа: http://ru.wikipedia.org/wiki/Apertium.

2. Виртуальная лаборатория Apertium [Электронный ресурс]: – Режим доступа: http://ru.vlab.wikia.com/wiki/Apertium.

ДО ПРОБЛЕМИ СТАНОВЛЕННЯ СИСТЕМИ УКРАЇНСЬКОЇ ТЕРМІНОЛОГІЇ ІНТЕЛЕКТУАЛЬНОЇ ВЛАСНОСТІ

Архипенко Л. М.

*Національний технічний університет
"Харківський політехнічний інститут",
м. Харків, вул. Пушкінська, 79/2*

Сьогодні ефективність інтелектуальної, творчої діяльності людини без перебільшень можна назвати поступом цивілізації. Головний чинник сталого соціально-економічного розвитку держави – зростання інтелектуального потенціалу нації, упровадження науково-технологічнх новацій, які суттєво впливають на обсяги та якість виробництва і споживання.

Характерною ознакою кінця XX – початку XXI століття стала масова поява на ринку нового виду товару – об'єктів права інтелектуальної власності. При цьому темпи росту обсягів торгівлі цим товаром зростають значно швидше, ніж для звичайних товарів. Згідно з прогнозними оцінками фахівців, інтелектуальна власність у цьому столітті стане основним товаром внаслідок переходу від індустріального до інформаційного суспільства. Розвиток ринку інтелектуальної власності зумовив необхідність відособлення цілого пласта лексичних одиниць в окрему терміносистему.

Актуальність дослідження пов'язана з необхідністю детальної характеристики терміносистеми «інтелектуальна власність» із урахуванням новітніх процесів, що сприяють її постійному поповненню.

У межах доповіді розглянемо формальні критерії дослідження термінології інтелектуальної власності (структурний склад термінів (співвідношення одно-, двох- і багатослівних термінів і терміносполучень);

морфологічні й синтаксичні деривaційні процеси (основні способи термінотворення)), оскільки структурні особливості й деривaційні процеси відіграють важливу роль для розуміння термінів, для взаєморозуміння спеціалістів галузі інтелектуальної власності. Аналізуючи формальну сторону терміна, більшість дослідників виділяють дві основні групи, які різняться своїм складом: терміни-слова (однокомпонентні) і терміни-словосполучення (багатокомпонентні). У досліджуваній терміносистемі присутні як однокомпонентні, так і багатокомпонентні терміни.

Проведені раніше системні дослідження різних термінологій свідчать про те, що у вже сформованій термінології тієї чи іншої галузі знань, кількість багатокомпонентних термінів становить близько 70%. За нашими приблизними підрахунками українська терміносистема «інтелектуальна власність» перебуває у стадії формування. Перевага багатокомпонентних термінів над однокомпонентними також свідчить про інтенсифікацію процесу спеціалізації термінів інтелектуальної власності, формування ієрархічної структури терміносистеми, встановлення деривaційних зв'язків між термінами.

Аналіз морфемної структури термінів інтелектуальної власності дозволяє виділити наступні типи однокомпонентних термінів: 1) прості непохідні терміни, основа яких збігається з коренем: *автор, запис, реферат, плагіат, райтер, роялті, аналог, патент, апеляція, домен, експерт, збір, знак, реєстр, особа, пріоритет, товар тощо;* 2) похідні терміни, тобто однослівні лексичні одиниці, основа яких має корінь і афікси.

Аналіз показує, що основним афіксом у термінотворенні є суфікс. До найбільш продуктивних суфіксів можна віднести: **-ання** (*анулювання, видання, використання, відтворення, копіювання, ліцензування*); **-ація** (*капіталізація, апеляція*); **-ник** (*винахідник, заявник*); **-ація** (*реєстрація, публікація)*; **-ор / -ер / -ар** та **-инг**, що свідчить про наявність у системі англомовних словотвірних елементів (*антрепренер, райтер, продюсер*).

Префіксація і префіксально-суфіксальний спосіб не виявляють високу продуктивність у творенні термінів інтелектуальної власності (*переробка, співавтор*).

До найбільш продуктивних способів творення термінології інтелектуальної власності належить словоскладання. Розглянемо найпродуктивніші структурні моделі багатокомпонентних термінів української терміносистеми «інтелектуальна власність»: іменник + іменник (*види порушень, використання товару, використання прав, депонування товару, опублікування твору*); прикметник + іменник (*авторський договір, авторські організації ефірне мовлення, ліцензійний договір, паушальний платіж*); іменник + прийменник + іменник (*випуск у світ, адреса для листування, відмова від свідоцтва, підходи до оцінки*); іменник + прикметник + іменник (*авторський сценарний договір, обмеження авторського права, виправлення очевидних помилок*); прикметник + прикметник + іменник (*безкоштовний обов´язковий примірник*); іменник + іменник + іменник + іменник (*найменування місця походження товару, продовження строку дії товару*); іменник + іменник + прикметник + іменник (*строк дії охоронного документа, роздільна здатність торгівельної марки, знак охорони*).

Проведений аналіз показав, що синтаксичне утворення термінологічних словосполучень є найбільш продуктивним способом термінотворення. При цьому у вибірці переважають двохкомпонентні сполучення, що забезпечує стрункість системи й перешкоджає її загромадженню більш довгими номінаціями. Як приклад розглянемо термінологічну групу з базовим терміном «*патент* – охоронний документ, що засвідчує право власника на деякі об´єкти промислової власності» (модель «прикметник + іменник): *ввізний патент* – патент, який оформлюється у спрощеному порядку на основі раніше виданого іноземного патенту; *деклараційний патент* – патент, який видається за результатами формальної експертизи заявки.

Отже, результати аналізу показують, що українська терміносистема «інтелектуальна власність» перебуває у стадії формування й розвивається й поповнюється за рахунок термінологічних сполучень та словоскладання, а зазначені способи термінотворення можуть зазнавати впливу аналогічної англомовної терміносистеми, що потребує подальших досліджень.

Список літератури

1. *Біла книга. Інтелектуальна власність в інноваційній економіці України* / [Г. О. Андрощук, О. В. Дем'яненко, І. Б. Жиляєв, та ін.] / упоряд. С. В. Таран. — К: Парламентське вид-во, 2008. — 448 с.

2. *Варфоломеева Ю.А.* Интеллектуальная собственность в условиях инновационного развития: Монография. – М.: «Ось-89», 2006. – 144 с.

3. *Інтелектуальна власність : Словник-довідник* / За заг. ред. О. Д. Святоцьког. – у 2-х т. : Том 1. авторське право і суміжні права / за ред. О. Д. Святоцького, Р. В. Дроб'язка [уклад : В. С. Дроб'язко, Р. В. Дроб'язко]. – К. : Видавничий дім «ІнЮре», 2000. – 356 с.

4. *Інтелектуальна власність* : Словник-довідник / За заг. ред. О. Д. Святоцьког. – у 2-х т. : Том 2. Промислова власність / За ред. О. Д. Святоцького, Р. В. Дроб'язка. [уклад : Г. П. Добриніна, А. В. Кочеткова, Н. І. Мова]. – К. : Видавничий дім «ІнЮре», 2000. – 272 с.

ДЛЯ НОТАТОК

ДЛЯ НОТАТОК